Spring 2017

Examining the type I error and power of 18 common post-hoc comparison tests

Derek Sauder
James Madison University

Examining the Type I Error and Power of 18 Common Post-hoc Comparison Tests

Derek Sauder

A thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Psychological Sciences, M. A.

Graduate Psychology

May 2017

FACULTY COMMITTEE:

Committee Chair: Christine DeMars, Ph.D

Committee Members:

Allison Ames Boykin, Ph.D

S. Jeanne Horst, Ph.D

Acknowledgments

First, I need to thank my advisor, Dr. Christine DeMars, for her patient and very much needed guidance on this project. Without her support and expertise, I never would have been able to complete this work. Similarly, my thesis committee was instrumental to success through their instructive feedback, so thank you Allison and Jeanne, as well.

Thanks to my quant cohort, Aaron and Catie. You guys helped keep me sane and thankful for the relatively limited scope of my project. The entirety of the Center for Assessment and Research Studies, too, deserves thanks. To both the faculty that listened to my project idea and the students that listened to me gripe about it, thank you!

Kierra, thank you. You may not have even known what my project was, but you were still supportive. Thank you for letting me pursue my master's degree, and thank you doubly for letting me continue on to my Ph.D. I promise it'll be your turn to go back to school soon!

Thanks to anyone else who deserves being thanked but that I didn't think of when writing this acknowledgment. You'll know who you are.

Table of Contents

List of Tables

List of Figures

Abstract

Researchers utilizing either experimental or quasi-experimental research often want to compare group means. However, with more than two groups, comparing group means may result in an inflated Type I error rate, the probability of wrongly rejecting a null hypothesis. Researchers often employ analysis of variance (ANOVA) methodology to compare more than two group means. Post-hoc comparison procedures (PCPs) are utilized to indicate which group means differ following a significant ANOVA. SPSS provides 18 options for PCPs. The purpose of this study was to determine which PCP provides the best power while maintaining Type I error control when assumptions of ANOVA are met and when they are not met. Data were simulated in a variety of conditions to address this issue. Only those tests designed for assumption violations, Dunnett's C, Dunnett's T3, Games-Howell, and Tamhane, adequately controlled Type I error in all conditions. Power results were similar for all four tests, with the Games-Howell being slightly higher than the other four tests. I recommend using the Games-Howell procedure unless extenuating circumstances exist.

Keywords: multiple comparisons, power, Type I error, ANOVA, post-hoc comparison procedures

Chapter 1. Purpose of Study

Random-assignment experimental research is a staple of fields such as business, psychology, and medicine (though terminology may differ between fields). Although experimental research takes different forms in each study, the fundamental principles are the same: comparisons between a randomized control group(s) and a randomized, systematically manipulated experimental group(s). An accessible example of experimental research is in a drug testing experiment, where one randomly assigned group is given a placebo pill and another randomly assigned group is given the drug of interest. Then, if groups are sufficiently randomized, any differences in the dependent variable under study can be attributed to the effects of the drug.

However, in certain situations, it is either impossible or infeasible to randomize participants to control and experimental groups. For example, perhaps the drug of interest is a dietary supplement designed to aid in weight loss. Further, the drug is expected to be more efficacious for overweight and morbidly obese individuals than for non-obese individuals. To test this claim, a researcher would need to collect samples from the populations of overweight, morbidly obese, and non-obese individuals. However, it is impossible to assign people to be overweight, obese, or non-obese. Consequently, the experimental (overweight and obese) and control (non-obese) groups are not randomly assigned. This type of research is often referred to as quasi-experimental.

In both experimental and quasi-experimental research, oftentimes the outcome (dependent variable) of interest is a group mean. For example, in a weight loss drug study, the dependent variable may be pounds lost for the overweight, obese, and non-obese participants. Consequently, research questions frequently center on comparing group means to one another (e.g., did obese individuals have higher average weight loss than non-obese or overweight individuals?). One popular way of comparing two group means is by an independent samples *t*-test. However, if there are more than two groups in a study, a researcher may wish to compare all three groups to each other in a pairwise manner. Thus, the researcher may utilize three independent samples *t*-tests comparing groups 1 and 2, groups 1 and 3, and groups 2 and 3. While appealing for its simplicity, this approach has an inherit risk of multiplicity.

Multiplicity

Whenever a researcher utilizes a statistical test, there is a risk of wrongly rejecting the null hypothesis due to sampling variability when the null is in fact true. The Type I error rate, denoted α, is the probability that the null hypothesis is falsely rejected and is commonly set at .05. Then, the probability of *not* making a Type I error is $1 - \alpha$, or .95 when $\alpha = .05$, as is typically chosen. When multiple tests are conducted on the same data, each individual test has an inherent Type I error rate of α. To determine the overall probability of making at least one Type I error over a set of independent tests, commonly called the familywise (Toothaker, 1993) or experimentwise (Ryan, 1959) Type I error, simply multiply the probabilities of not making a Type I error for each test together and subtract from 1. If $\alpha = .05$, this formula appears as $1 - (1 - \alpha)^m$ or $1 - (.95)^m$, where *m* is the number of tests being conducted (Field, 2013, chapter 2). Because .95 is the probability of not making a Type I error, we raise it to the power of *m* due to the multiplicative nature of probabilities (i.e., the probability of *m* number of independent events all occurring is equal to the product of their individual probabilities). This value, then, is the probability of not making any Type I errors. Next, we subtract from 1, because we are interested in the probability of making at least one Type I error. For the weight loss drug example, there would be three tests, and the familywise Type I error would be equal to $1 - (.95)^3 = .14$. Thus, there would be a 14% chance of falsely rejecting at least one null hypothesis *for the set of tests*. The more tests computed, the larger the familywise Type I error becomes. Obviously, this is an undesirable effect and is referred to as the multiplicity issue.

An astute researcher may simply ask, “If Type I error increases with the number of tests I conduct, why can’t I adjust my α level to account for such?” One popular procedure, called the Bonferroni procedure, does exactly that. Using the Bonferroni procedure in the three comparison example with the weight loss drug results in setting the alpha level at roughly .017 (.05/3), a much more conservative value than the typical .05. The trade-off to combating multiplicity by decreasing α is a loss of statistical power (Field, 2013, chapter 2). Power refers to the ability of a test to correctly identify a mean difference in a population of interest. Continuing with the weight loss example, the researcher may want a test with high power to detect if there are true differences between obese and non-obese participants, because the financial future of the drug depends on the results. By controlling Type I error by making α stricter, the test loses power because it allows for so few falsely rejected null hypotheses. Consequently, some hypotheses that should be truly rejected are not rejected.

Analysis of variance

Researchers often use analysis of variance (ANOVA) to test the null hypothesis of equal means for multiple groups simultaneously. The simplest ANOVA model is often called a one-way ANOVA, which consists of only one grouping independent variable with three or more levels and one continuous dependent variable. However, ANOVA can be utilized for a variety of research designs including repeated-measures designs, multiple factor designs (often called factorial ANOVA), or even multivariate designs (MANOVA). The popularity of ANOVA is widespread, with a recent study showing that it was taught at least every two years in 95% of doctoral psychology programs (Aiken, West, & Millsap, 2008). Elmore and Woehlke (1998) found that ANOVA and/or analysis of covariance (ANCOVA) type methodology was the second most employed method (behind descriptive analysis) across three journals (*American Educational Research Journal*, *Educational Researcher,* and *Review of Educational Research*) from 1978 to 1997. A similar study found factorial ANOVA to be the most common methodology in the *Journal of Educational Psychology* (Goodwin & Goodwin, 1985).

Without going into detail, researchers use ANOVA to partition group variance on the dependent variable into variance attributable to differences between groups (often called between-group variability) and within groups (often called within-groups variability or error variability). A ratio of these two types of variability is created to produce an *F*-statistic with a known distribution. By comparing the *F*-statistic to a critical value determined by α and degrees of freedom, ANOVA indicates whether group means differ statistically significantly from one another. If the *F*-statistic is significant, at least one of the group means differs from another[1].

Although ANOVA is a useful tool for comparing group means, it does lack the ability to specify *which* means differ following a statistically significant *F*-test. Knowing that one group differs from another may be helpful in a limited sense, particularly for a

[1] In certain situations, means may not statistically significantly differ if the omnibus *F*-test is significant due to a lack of power. See Field (2013, chapter 11) for more details about the inner workings of ANOVA.

small number of groups and largely disparate means, but when the number of groups increases and means are relatively close, more information is required. There are two primary ways of determining which means differ from each other: planned comparisons and post-hoc comparison procedures (PCPs).

Planned comparisons

Although not the focus of this study, a brief introduction to planned comparisons is warranted. Planned comparisons are a popular tool for comparing group means, particularly in confirmatory studies where groups are theorized to relate to one another in a certain way. Established *a priori*, planned comparisons allow a researcher to choose which group comparisons he or she desires to compute. Thus, researchers can test specific hypotheses (see Ruxton & Beauchamp, 2008). For example, in typical experimental research, one or more experimental groups are often compared to a control group. A planned comparison of the difference between the average of the experimental groups and control group could be conducted by inputting a series of dummy codes and weights into a regression analysis. For a more thorough explanation of how planned comparisons work, see Field (2013, chapter 11).

Post-hoc comparison procedures (PCPs)

PCPs, most of which were developed sometime during the 1950's to the 1980's, are more exploratory in nature than planned comparisons and are computed *after* running the ANOVA (hence, post-hoc). Most PCPs operate by comparing every group mean to every other group mean (referred to as pairwise comparisons). For example, in a three-group scenario, there would be three unique comparisons: the mean of group 1 to the mean of group 2, the mean of group 1 to the mean of group 3, and the mean of group 2 to the mean of group 3. In a four-group scenario, there would be six unique comparisons; for five groups, there are ten unique comparisons, and so on. Even though this sounds exactly like conducting multiple *t*-tests, which was shown to be poor practice, PCPs were all developed to account for the increased Type I error rate of conducting multiple tests in some way.

For example, one PCP, Fisher's least significant difference (LSD), was designed to control familywise Type I error by requiring a statistically significant ANOVA *F*-test prior to being computed (Fisher, 1935). Thus, Fisher's LSD would only be computed in error (i.e., if there are no significant differences to be found) if the significant result of the ANOVA *F*-test was itself a Type I error. Similar logic can be applied to all PCPs: if the omnibus *F*-test is not statistically significant, why would a researcher follow-up with a post-hoc test? However, all of these procedures (with the exceptions of Fisher's LSD) were designed as stand-alone procedures, not as follow-ups. Further, some statistical software will compute PCPs regardless of the results from the omnibus *F*-test. Consequently, examining the Type I error of these tests even when group means truly do not differ and the omnibus test is not statistically significant is still extremely valuable, particularly due to the frequent use of PCPs.

Partly because of the popularity of ANOVA, PCPs are also a popular choice among researchers. For example, Goodwin and Goodwin (1985) found that for a sample of 150 articles across a five-year span in the *Journal of Educational Psychology*, almost a third (47 out of 150) of articles employed some sort of PCP (referred to as "Post-hoc Multiple Comparisons" in the article). Similarly, Keselman et al. (1998) noted that 29 out of 61 articles (across a wide variety of psychology and education journals) that utilized between-subjects ANOVA designs also incorporated PCPs. To conduct ANOVAs and PCPs, many researchers turn to IBM's Statistical Package for the Social Sciences (SPSS). The availability of pull-down menus in SPSS renders unnecessary the knowledge of the syntax necessary to compute such tests in similar statistical programs such as R or SAS. Consequently, SPSS may be more intuitive and less intimidating to use for both seasoned and new researchers than R or SAS. Muenchen (2016) found that for scholarly articles, SPSS was used in approximately twice as many articles as the next two closest competitors (R and SAS, respectively).

However, one disadvantage of the pull-down menus is that researchers do not necessarily change the settings off default options, which are not always ideal for a given research question or dataset. Similarly, for certain procedures, there are many options to choose from, which can be both overwhelming and confusing. For example, SPSS 23 and 24 provide 18 different PCP options with little to no explanation of what each does or how it works. Indeed, Games (1971) noted, "The area of multiple comparisons is one of the more confusing areas of statistics, and is one that receives a widely differing set of recommendations from many applied statistics texts in behavioral sciences" (p. 531). Consequently, a researcher may inadvertently choose a PCP that does not perform well under certain conditions. Similarly, the number of choices may be so overwhelming that a researcher chooses one at random or elects not to use PCPs at all. The problem is compounded by violating the assumptions associated with conducting an ANOVA, because the nature of the violations further influences the choice of PCP.

Assumptions of ANOVA

The data assumptions associated with conducting an ANOVA come from the method of estimation used to determine the parameters of interest (in ANOVA, the *F*-statistic). Typically, ANOVA is estimated with ordinary least squares (OLS). OLS, as the name implies, produces parameter estimates that minimize the sum of the squared residual terms between the actual and method-implied data. OLS is a closed-form estimator, meaning that there is only one set of analytically derived parameters for a given procedure (in contrast, maximum likelihood [ML] estimation is an iterative process that only arrives at a solution when a certain criterion is met).

OLS has several assumptions that can be lumped into three categories: 1) the model is correctly specified, 2) there is no measurement error in the independent variables, and 3) the residuals are independent and identically normally distributed with a mean of 0 (Cohen, 2013, chapter 10; Pedhazur, 1997). In simpler terms, the first assumption means that the relationship between the independent variables (IVs) and dependent variables (DVs) is linear and that all relevant IVs are included. The second assumption has to do with reliability of measurement in that IVs are assumed to have a reliability of 1.0. The third assumption contains several pieces: residuals are uncorrelated with one another, normally distributed about 0, and have equal variances across groups (also called homoscedasticity or homogeneity of variances).

The first two assumptions of OLS (correctly specified model and no measurement error in IVs) and the assumption of independence of residuals are largely a concern during research design. Fortunately, because the IV in one-way ANOVA designs is a grouping variable, the IV should have high reliability. As for normality, OLS has been shown to be robust to violations of normality (Bohrnstedt & Carter, 1971; Boneau, 1960; Pedhazur, 1997). Indeed, "…non-normality has only minor consequences in situations represented by most research applications" (Hopkins & Weeks, 1990). Therefore, I focus solely on violations of the final assumption: homoscedasticity. Finally, ANOVA is typically conducted with equal group sizes due to the pooling of variance across groups. In reality, due to restrictions of sampling or various other reasons, this is often not the case. Fortunately, this issue has largely been resolved for ANOVA by weighting variances by group size. However, many of the older PCPs do not account for the possibility of unequal group sizes, and thus may not function appropriately when group sizes are unequal.

Study purpose

Let us return to the example of the weight loss drug. Recall that the researcher wanted to compare the weight loss of three groups: obese, overweight, and non-obese. Say, for this hypothetical example, that the assumptions of OLS are met except for homoscedasticity. Further, due to the relative minority of obese individuals, assume the group sizes are unequal, as well. The researcher conducts an ANOVA and attains a statistically significant *F*-statistic. However, the researcher does not know which PCP to choose for the data. SPSS does indicate that 4 of the 18 options are designed for unequal variances, but which of those four is "the best?" The researcher wants to maximize power while maintaining Type I error control. The purpose of this study is to explore the question: which PCP should I choose, given my data?

Chapter 2. Literature Review

Type I error

There are two main approaches to examining Type I error when conducting multiple comparison tests: familywise (sometimes called experiment-wise) and comparison-wise. Consider a scenario where teachers apply four different teaching styles for a semester long course. The outcome variable is final exam grade. Following a significant ANOVA, the researcher wants to conduct post-hoc comparisons. Controlling Type I error in a comparison-wise fashion means that the alpha level for each comparison of a pair of means is set to some nominal level (usually .05). This is analogous to conducting six *t*-tests in the example and is poor practice because the probability of making at least one Type I error among the six tests is greater than .05. By controlling familywise Type I error, the alpha level remains at the nominal level (or below) for a set (i.e., family) of comparisons. Thus, the overall alpha for the six comparisons of the example would remain at the nominal level. Due to the relative disadvantages of controlling Type I error comparison-wise and the popularity of familywise Type I error control, I focus on familywise Type I error. A procedure is said to control familywise error in the weak sense if it does so only when all null hypotheses are true and is said to control error in the strong sense if it does so for any configuration of true and false hypotheses (Benjamini & Hochberg, 1995; Hochberg, 1988).

False discovery rate

Instead of only considering wrongly rejected null hypotheses, the total number of null hypothesis rejections can also be examined. Benjamini and Hochberg (1995) suggested an alternate way of defining the issue of multiple comparisons: the false discovery rate (FDR; see also Curran-Everett, 2000). The FDR is the proportion of wrongly rejected null hypotheses divided by the total number of rejected null hypotheses. The major advantage to the FDR is the increase in power over controlling for familywise Type I error due to the ability to set an acceptable level of false rejection. Thus, procedures controlling the FDR will be more likely to find true differences when compared with familywise error control, particularly when there are more true differences to find. Further, Benjamini and Hochberg (1995) showed that controlling the FDR also controls familywise error in the weak sense. For an example of how a test may provide control over FDR, see Keselman, Cribbie, and Holland (1999). However, because SPSS utilizes procedures designed to control familywise error, final recommendations will give more weight to familywise Type I error rate and less to FDR.

Power

If differences in population means truly exist, Type I errors cannot be committed because the null hypothesis is false. Thus, statistical power of tests must also be considered. Power in mean comparison tests is traditionally conceptualized as the probability that the researcher rejects a null hypothesis based upon the test when there are true differences[2] and is equal to $1 - \beta$ where β is Type II error[3] (Field, 2013, p. 69). In other words, power is the ability to detect a difference in the population when there is one. For example, power for an ANOVA is the probability that the *F*-ratio will be statistically significant *if* there truly exists at least one difference among a set of means. Power can be conceptualized in several different ways, however, for PCPs. Specifically, one can examine per-pair power, any-pair power, or all-pairs power (Demirhan, Dolgun,

[2] More generally, power is the probability of rejecting a null hypothesis if the null is false.
[3] A Type II error occurs when true population differences are undetected by a statistical test.

Parlak, & Dolgun, 2010; Jaccard, Becker, & Wood, 1984). Per-pair power is exactly as it sounds: the power for a given comparison. Any-pair power refers to the probability of correctly rejecting at least one null hypothesis for a set of comparisons. Thus, any-pair power is analogous to familywise error. Last, all-pairs power is the probability that all false null hypotheses are rejected. Obviously, all-pairs power is a far stricter measure than any-pair power in most cases. I will report all types of powers, but give more weight to any-pair power when making PCP recommendations due to its similarity to familywise error and the strictness of all-pairs power. However, power (and Type I errors) can only occur under certain distributions.

Null and alternative distributions

Defining null and alternative distributions in the context of PCPs is conceptually easiest when examining mean differences between groups. Under the null distribution, the null hypothesis is that the mean difference is equal to zero. Thus, the mean of one sampled group is equal to another sampled group, within sampling error, because both groups come from the same population. However, under the alternative distribution, the alternative hypothesis states that the mean difference is not equal to zero, and the two sampled groups must come from different populations.

A Type I error can only be committed when the null distribution is true and a Type II error (1 – power) can only be committed when the alternative distribution is true. It is impossible to commit a Type I error if there are true population differences, as is the case when the alternative distribution is true, because the null hypothesis should be rejected. Similar logic applies to Type II errors: only when the alternative distribution is true and the null hypothesis should be rejected, but is not, can an error occur.

In real data situations, researchers cannot know which distribution, the null or alternative, is the “truth.” Thus, researchers employ statistical tests (e.g., PCPs) to determine if the data (e.g., group means) are from the same (null distribution) or different (alternative distribution) populations. Statistical tests set a nominal Type I error rate (i.e., α/false positives) and attempt to maximize power (i.e., minimize Type II errors/misses of statistical significance). In real data situations, it is impossible to know if a correct decision or an error (Type I or Type II) is being committed. Thus, researchers must trust that Type I errors occur at the nominal level and maximize power by increasing sample size, making the treatment effect stronger, or reducing the mean square error (MSE) with statistical controls. Fortunately, in simulation studies, researchers know if the data were simulated to follow the null or the alternative distribution. As such, empirical Type I error and power rates can be computed for various data conditions (such as when assumptions of estimators are violated) to determine how often statistical tests result in errors. Let us now turn our attention to the statistical tests in question: the PCPs.

Simultaneous versus sequential procedures

A brief explanation of two classes of PCPs is required: simultaneous and sequential (Toothaker, 1993). Simultaneous test procedures (STPs) control for the Type I error for a set of comparisons and use one alpha value for all comparisons. STPs include tests such as Tukey’s honestly significant difference (HSD) and Scheffé tests. Sequential (also called stepwise) procedures utilize a series of comparative steps. The test only proceeds to the next step if the one before it meets certain criteria (e.g., statistical significance). Most sequential procedures use step-down logic, where the largest difference in means is tested first before moving on to the next largest mean difference, and so on. Examples of sequential procedures using step-down logic are the Student-Newman-Keuls (SNK) and the Ryan-Einot-Gabriel-Welsch Q (REGWQ). In contrast, some sequential procedures use step-up logic, where test statistics ordered from smallest to largest are compared to critical values (Dunnett & Tamhane, 1992; Hochberg, 1988). If the test statistics are significant (i.e., larger than the critical value), any larger test statistics are also deemed significant (see footnote 23 in Toothaker, 1993). For example, in a set of ordered test statistics, if the first statistic (derived from means 1 and 2) is statistically significantly, then the statistic for means 1 and 3, 1 and 4, etc. are also statistically significant. None of the PCPs in SPSS utilize a step-up procedure.

Description of PCPs

This section details the 18 PCPs that are available in SPSS 23 and 24. There are many additional PCPs available, and interested readers should consult Keselman, Cribbie, and Holland (2004) or Klockars and Hancock (1992) for some (relatively) newer procedures. However, because of the popularity of SPSS, only the 18 available PCPs are examined. Each PCP is briefly described and then followed by the SPSS algorithm used to compute the PCP. The formulas utilized by SPSS may differ from what the original authors described, but, because of my interest in studying the way SPSS computes PCPs, I used the SPSS formulae (IBM, 2014) instead of the original formulae. Before beginning, notation used by SPSS is detailed in Table 1, which is a recreation of the information found in Appendix G of the SPSS 23 algorithm guide (IBM, 2014).

SPSS computes some of the 18 PCPs using one of two range statistics. The more common range statistic is a Studentized range value. The Studentized range, traditionally denoted as *q*, is similar to the *t*-statistic and is equal to the difference between the largest and smallest mean over the square root of the mean square error (MSE) divided by *n* (IBM, 2014; Winer, 1971). SPSS denotes the Studentized range value as $S_{\varepsilon,r,m}$, where ε is equal to (1 – α) for a one-tailed test and (1 – α)/2 for a two-tailed test, *r* is the total number of means being compared, and *m* is a measure of degrees of freedom. The ε, *r,* and *m* variables may differ for the 18 PCPs.

The second range statistic is the Studentized maximum modulus. Similar to the Studentized range, the Studentized maximum modulus is equal to the maximum of the absolute values of the group means divided by an estimate of the sample standard deviation with *m* degrees of freedom (IBM, 2014; Stoline & Ury, 1979). SPSS denotes the Studentized maximum modulus as $M_{\varepsilon,r,m}$, where ε, *r*, and *m* are defined the same as for the Studentized range. Again, the values for the ε, *r*, and *m* variables may change depending on the PCP. Finally, several tests use neither the Studentized range value nor the Studentized maximum modulus. For these tests, the full formula is given as is detailed in IBM (2014).

For most PCPs, SPSS outputs a table of pairwise comparisons. This table contains mean differences, standard errors, significance (i.e., *p*-values), and confidence intervals for each possible pairwise comparison. For certain tests, which I will highlight below, SPSS also outputs information on homogeneous subsets. Means are placed in a homogeneous subset if they are not statistically significantly different. The maximum number of homogenous subsets is therefore equal to the number of groups.

Information from existing simulation studies will be described in the final paragraph for each PCP. The majority of studies examined tests under violation of assumptions. However, although some simulation studies examined both Type I error and power, most only focus on one or the other. Consequently, there may be little to no information about how tests perform (in terms of power and Type I error) under violation of assumptions. Throughout this section, the terms "conservative" and "liberal" refer to Type I error rate whereas the terms "increased/high" or "decreased/low" will refer to power.

PCPs for equal variances. The following PCPs were designed for data that are homoscedastic. Unfortunately, this is rarely the case in real data. However, most PCPs are robust to some deviation from homoscedasticity. One popular way of testing the homogeneity of variances assumption in SPSS is the Levene's test. If Levene's test is statistically significant, the assumption of homogeneity of variances is violated. However, Levene's test is influenced by sample size because it is a null hypothesis statistical significance test and will always be significant for real data with a large enough sample. Visual inspection of residuals and examination of the ratio of largest to smallest variance (sometimes called F_{max}) are additional methods for determining if homoscedasticity is violated to a practical extent.

Fisher's least significant difference (LSD). Fisher's LSD was the first PCP created (Fisher, 1935). This method of comparison actually has no form of Type I error control beyond the assumption that the omnibus ANOVA test is significant[4]. However, SPSS will compute the LSD regardless of the overall ANOVA *F*-test. The LSD is analogous to computing a series of *t*-tests on a set of means. The only difference is that

[4] The requirement of a significant *F*-statistic does maintain the total proportion of times where one or more PCPs is falsely rejected at the nominal alpha. However, the total proportion of falsely rejected PCPs will be greater than the nominal alpha due to dependence among PCPs in an experiment. Additionally, if the null hypothesis is partly true, Fisher's LSD does nothing to control the Type I error rate for the comparisons that *do* have true null hypotheses.

the standard deviation is a pooled standard deviation across all group means instead of a pooled standard deviation of the two means being compared. Thus, we expect familywise Type I error to be $1 - (1 - \alpha)^{k^*}$, where k* is the number of comparisons made. A comparison between two means is significant if the following is true:

$$\overline{x_i} - \overline{x_j} \ > Q_{i,j}\sqrt{2F_{1-\alpha}(1, f)}, \tag{1}$$

where $\overline{x_i}$ and $\overline{x_j}$ are the means for groups *i* and *j*, respectively, $Q_{i,j}$ is equal to $s_{pp}\sqrt{\frac{1}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$ where s_{pp} is the square root of MSE term from the omnibus ANOVA *F*-test and n_i and n_j are group sizes for *i* and *j*, respectively, and $F_{1-\alpha}(1, f)$ is the critical value of the *F*-distribution with degrees of freedom equal to 1 and the degrees of freedom for the MSE term (*f*).

Conducting multiple *t*-tests on the same data inflates Type I error, which is only exacerbated by violating assumptions, because heterogeneous variances also inflate Type I error of *t*-tests (Boneau, 1960). Thus, the LSD will not maintain Type I error control when assumptions are met or otherwise. However, because the LSD will often be statistically significant, the test does offer the researcher high power (i.e., the more null hypotheses rejected, the more likely to correctly reject one).

Bonferroni. The Bonferroni procedure was popularized by Dunn (1959; 1961). Named for its use of Bonferroni inequalities, the Bonferroni method controls for Type I error by adjusting the alpha level for each pairwise comparison. In fact, the formula for each comparison is the same as the LSD in equation 1 except for α. Instead of being set at the nominal .05 level, α is computed as follows:

$$\alpha' \ = \ \epsilon/k^*, \tag{2}$$

where ϵ is equal to the nominal familywise error rate. Thus, the alpha level for each pairwise comparison is equal to the desired familywise error divided by the number of comparisons being made. This is a direct attempt to combat the multiplicative increase in Type I error for conducting multiple comparisons.

Dunn (1961) noted that when the number of comparisons is large, the Bonferroni method results in wider confidence intervals (i.e., less power) than other methods. Relative to other PCP methods, the Bonferroni procedure also may not detect group mean differences even when assumptions are met (Curran-Everett, 2000). Moreover, there is evidence that unequal group sizes increase Bonferroni Type I error rates, whereas heterogeneous variances have little effect (Demirhan et al., 2009). Kromrey and La Rocca (1995) concluded that the Bonferroni procedure (which they refer to as the Dunn procedure) maintained Type I error control in a liberal sense (i.e., less than .075 when nominal alpha was .05) for unequal variances in most cases. However, Type I error was still inflated.

Sidak. The Sidak (1967) test is a modification of the Bonferroni procedure that provides slightly more power by allowing a slightly larger α for each comparison. Instead of using equation 2 to modify the alpha level used with equation 1, the following equation is used:

$$\alpha'' = 1 - (1 - \epsilon)^{\frac{1}{k^*}}. \tag{3}$$

The above equation is derived by solving equation 4 for α_c, the per comparison error rate for *m* multiple comparisons if the nominal familywise error rate is α_f:

$$\alpha_f = 1 - (1 - \alpha_c)^{m}. \tag{4}$$

Thus, the Sidak equation actually determines the precise alpha level per comparison (α_c) to ensure the overall familywise error rate (α_f) is the nominal level (usually .05). The Bonferroni procedure corrects for Type I error in a strict additive sense (the alpha per comparison sums to the desired familywise alpha) which is more conservative than necessary because the familywise error is not equal to the sum of the error rates for each comparison. The Sidak method provides a more exact alpha level per comparison so that the familywise error will be the same as the nominal level (instead of below it), thus increasing power. However, the Sidak test loses power when group sizes or variances are unequal (Demirhan et al., 2009).

Student-Newman-Keuls (SNK). The Student-Newman-Keuls (SNK) procedure is, unsurprisingly, named for three papers by Student (1927), Newman (1939), and Keuls (1952). This test is a sequential PCP that uses the step-down procedure to compare means. Thus, means are ordered and the largest and smallest are compared first. If the largest and smallest means statistically significantly differ, the next smallest is compared with the largest and the smallest is compared with the next largest, and so on until all comparisons are made. Comparisons that are not statistically significant are placed in homogeneous subsets, which SPSS displays in the output. If two group means are in different subsets, they differ statistically. A comparison is statistically significant if the following equation holds:

$$\overline{x_i} - \overline{x_j} \; > Q_h S_{\varepsilon,r,f}, \tag{5}$$

where ε is equal to (1 – α)/2 for a two-tailed test, *r* is the number of steps between the ordered means being compared, *f* is the degrees of freedom from the MSE term, and Q_h is equal to $s_{pp}/\sqrt{n_h}$ where n_h is the harmonic mean of the sample size, $\frac{k}{\sum_{1\leq i\leq k} n_i{}^{-1}}$. This is the first test described that utilizes the Studentized range statistic, which is similar to a *t*-statistic. Note that the critical value of the Studentized range statistic (and consequently the SNK test) depends on *r*, the number of steps between means, and thus differs across comparisons. Thus, the critical value for each comparison will depend on how close the means being compared are relative to all the other group means.

The SNK does not control Type I error when assumptions are met except in the special case of three groups (Einot & Gabriel, 1975; Ramsey, 1978; Ramsey, 1981). Further, when assumptions are not met, the SNK is negatively affected by unequal group sizes and variances in terms of both power and Type I error (Demirhan et al., 2009), particularly when the smallest group has the largest variance (Petrinovich & Hadrych, 1969) and as the number of groups increases (Kromrey & La Rocca, 1995).

Tukey's honestly significant difference (HSD). Tukey's honestly significant difference (HSD) test (also called Tukey's A and, sometimes, wholly significant difference [WSD]) is one of the most popular PCPs used (if not the most popular). Described as what "may be the most frequently cited unpublished paper in the history of statistics" (Toothaker, 1993, pp. 32-33), Tukey first introduced the HSD in a mimeographed monograph. This procedure, similar to the SNK, also utilizes the Studentized range statistic. In fact, Tukey's HSD is computed in the same way as in equation 5, except for the two differences seen in equation 6. First, instead of using *r* to compute the critical value, Tukey proposed using *k*, the number of groups. Thus, the critical value for each comparison is the same, because *k* is constant, meaning all comparisons are computed simultaneously. Second, Tukey's HSD uses $Q_{i,j}$ instead of Q_h:

$$\overline{x_i} - \overline{x_j} \ > Q_{i,j} S_{\varepsilon,k,f}. \qquad (6)$$

SPSS will produce redundant homogeneous subset output and pairwise comparison output for Tukey's HSD. Tukey's HSD was designed for equal variance and equal sample sizes. If these assumptions are violated, the test can become either more conservative or too liberal.

When assumptions are met, Tukey's HSD controls Type I error well with power that is about average, being greater than some and less than other PCPs (Petrinovich & Hadrych, 1969; Ramsey, 1981; Seaman, Levin, & Serlin, 1991). However, when assumptions are not met, Tukey's HSD does not strictly control Type I error when variances are unequal at ratios of 13:1 (Kromrey & La Rocca, 1995) or the smallest group has the largest variance (Petrinovich & Hardyck, 1969).

Tukey's B. Far less well known is Tukey's B (which is unfortunately also frequently referred to as the wholly significant difference [WSD] test and thus sometimes confused with Tukey's HSD). Tukey's B is a compromise between the SNK and Tukey's HSD tests. The range statistic is computed as the average of the Studentized range statistics from the two tests: ½($S_{\varepsilon,r,f}$ + $S_{\varepsilon,k,f}$). Thus, as with the SNK, each comparison will have a slightly different critical value associated with it. Additionally, Tukey's B uses the harmonic mean (Q_h), like the SNK. SPSS outputs homogeneous subset information for Tukey's B.

Because Tukey's B is a compromise between the SNK and the HSD, it will perform somewhere in the middle in terms of Type I error control and power when assumptions are met, and will control Type I error adequately for three groups (Petrinovich & Hardyck, 1969). In other words, Tukey's B will be more conservative than the SNK but more powerful than the HSD (Duncan, 1955; Petrinovich & Hardyck, 1969). When assumptions were violated, Tukey's B was too liberal with three groups if the smallest group had the largest variance (Petrivonich & Hardyck, 1969).

Scheffé. Scheffé developed his method for simultaneously computing all possible comparisons (not just pairwise comparisons; Scheffé, 1953). The advantage to Scheffé's method is that it allows a researcher to conduct any post-hoc comparison he or she desires. The trade-off for this is lower power and a too conservative Type I error rate. Further, researchers often only care about pairwise comparisons, making the utility of the Scheffé test a moot point. A given comparison is statistically significant if:

$$\overline{x_i} - \overline{x_j} \ > Q_{i,j}\sqrt{2(k-1)F_{1-\alpha}(k-1,f)}. \tag{7}$$

SPSS provides both homogeneous subset output and pairwise comparison data for the Scheffé test.

The Scheffé test tends to be conservative and underpowered when data assumptions are met and the number of groups is large (Games, 1971; Ozkaya & Ercan, 2012; Petrinovich & Hardyck, 1969). Violations of assumptions can serve to exacerbate or lessen this problem depending on the manner of the violations (Keselman & Rogan, 1978; Petrivonich & Hardyck, 1969).

Duncan's multiple range test. Duncan's (1955) multiple range test (MRT) is very similar to the SNK, but designed with an increase in power in mind. The difference between the two tests lies in an adjustment to the Studentized range value's alpha level. Instead of using the nominal familywise error (ϵ), Duncan's test uses the following formula:

$$\alpha \ = 1 - (1 - \epsilon)^{\, r-1}. \tag{8}$$

The change to the alpha level used in computing the Studentized range value results in more liberal tests for those comparisons where the range between means is larger.

When assumptions are met, the multiple range test tends to inflate Type I error rate (Carmer & Swanson, 1973), particularly when $k > 3$ (Petrinovich & Hardyck, 1969; Seaman, Levin, & Serlin, 1991). Type I error for Duncan’s multiple range test is affected by unequal group sizes (Demirhan et al., 2009). Ozkaya and Ercan (2012) also found inflated Type I error rates when group sizes differed. Unequal variances inflate Type I error if the smallest group has the largest variance (Petrinovich & Hardyck, 1969).

Hochberg’s GT2. The Generalized *T* procedures (GT1 and GT2) were originally designed as a way of extending Tukey’s HSD for data with non-homogenous variances or unequal covariances (Hochberg, 1974). The GT2 was shown to provide more power than the Bonferroni and Scheffé tests when assumptions were met due to its use of the Studentized maximum modulus. A pairwise comparison is statistically significant if the following inequality holds:

$$\overline{x_i} - \overline{x_j} \ > Q_{i,j}\sqrt{2}M_{\varepsilon,k^*f}, \tag{9}$$

where M_{ε,k^*f} is the Studentized maximum modulus with degrees of freedom equal to k^*, the number of comparisons, and *f*, the degrees of freedom for the MSE term. SPSS will give both homogeneous subset output and pairwise comparison output for Hochberg’s GT2.

The GT2 can be conservative when sample sizes and/or variances are unequal (Demirhan et al., 2009; Dunnett, 1980a). Conversely, the GT2 has also been shown to have largely inflated Type I error rates when the smallest group has the largest variance (Keselman, Games, & Rogan, 1979), a condition that usually causes inflated Type I error

(Glass, Peckham, & Sanders, 1972). Still other research has found that the GT2 was robust to assumption violations, except when the variances were unequal at larger proportions (such as 1:10; Tamhane, 1979).

Gabriel. Gabriel's (1978) pairwise comparison test was designed for comparison of confidence intervals when group sizes differed. Two means were said to be statistically significantly different if and only if their respective confidence intervals computed via Gabriel's method were disjoint. The algorithm utilized by SPSS for pairwise comparisons is:

$$\left|\overline{x_i} - \overline{x_j}\right| \geq s_{pp}\left(\frac{1}{\sqrt{2n_i}} + \frac{1}{\sqrt{2n_j}}\right) M_{\varepsilon,k^*,f}. \qquad (10)$$

The Gabriel test in SPSS will provide both homogeneous subset output and pairwise comparison output. When computing the homogeneous subsets, the harmonic mean n_h is used instead of n_i and n_j. Thus, slightly different results could arise when comparing the pairwise and homogeneous subset output.

Gabriel (1978) found that largely imbalanced group sizes tended to result in inflated Type I error, though the test was conservative for less disparate sample sizes. Dunnett (1980a) found that Gabriel's test was conservative when group sizes were unequal but variances were equal. If the differences among group sizes became too large, Gabriel's test became too liberal. Demirhan et al. (2009) also found that unequal group sizes increased Type I errors. Further, they found that heterogeneous variances affected Type I errors, and Keselman, Games, and Rogan (1979) found that when the smallest group size was paired with the largest variance, Type I error increased dramatically.

Waller-Duncan t-test. The Waller-Duncan *t*-test works similarly to Fisher's LSD, but instead employs Bayesian methods to create homogeneous subsets (Duncan, 1965;

Waller & Duncan, 1969). Additionally, the Bayesian *t*-statistic (t_B) is based in part on a relative seriousness ratio of Type I to Type II error (*w*)[5]. The equation used for conducting the test is:

$$v_{i,j} = \overline{x_i} - \overline{x_j} \geq t_B(w, F, q, f) s_{pp} \sqrt{\frac{2}{n}}, \qquad (11)$$

where *F* is the *F*-ratio from the one-way ANOVA, $q = k - 1$, $f = k(n - 1)$, and *n* is the group size. The default *w* ratio in SPSS is set at 100:1, which approximates to an alpha level of .05 (alternately, a ratio of 50:1 approximates an alpha of .10 and a ratio of 500:1 approximates an alpha of .01). As the *F*-ratio increases, t_B decreases, resulting in a more powerful test when assumptions are met. Equation 11 is for equal sample sizes; if sample sizes are unequal, n_h is used in place of *n*.

Waller and Duncan (1969) noted that their Bayesian *t*-test tends to inflate Type I error rate when the accompanying *F*-test is moderate to large, but is more conservative when *F* is small. Similarly, Carmer and Swanson (1973) noted an inflated Type I error rate, particularly as the number of comparisons increased. However, they noted that the Waller-Duncan test had good power.

Dunnett's t-tests. Dunnett (1955) proposed a special solution to the multiple comparison problem when a researcher wishes to compare treatment groups to a control group. Because of the nature of this design, Dunnett showed that the confidence intervals constructed around the means were narrower than those created by Tukey's HSD or Scheffé's test, thus increasing power when assumptions were met. Dunnett provided equations for two-tailed or one-tailed tests against the control group. SPSS also offers these capabilities, but I will limit myself to the two-tailed case because this is the more

[5] That is, *w* is a user-defined ratio of which error is considered more detrimental: Type I or Type II.

popular and conservative test. To compute Dunnett's two-tailed *t*-test, see equations 12.1 to 12.3[6].

$$|\overline{x_i} - \overline{x_0}| \ > d^{\varepsilon}_{k,v} s_{dd}\sqrt{\tfrac{1}{n_0}+\tfrac{1}{n_i}}, \tag{12.1}$$

where x_0 is the control group and $d^{\varepsilon}_{k,v}$ is the upper 100ε percentage point of the distribution of:

$$T = \max_{1 \le i \le k}\{|T_i|\}\ , \tag{12.2}$$

where

$$T_i = \frac{(\overline{x_i} - \overline{x_0})}{s_{dd}\sqrt{\frac{1}{n_0}+\frac{1}{n_i}}} \text{ and } s^2_{dd} = \frac{\sum_{i=0}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x_i})^2}{\sum_{i=0}^{k}(n_i - 1)}. \tag{12.3}$$

Dunnett's *t*-tests were designed for equal groups and will only provide approximate values when group sizes differ (Dunnett, 1955).

Ryan-Einot-Gabriel-Welsch (REGW) tests. There are two tests that arose out of a series of papers by Ryan (1960), Einot and Gabriel (1975), and Welsch (1977): the Ryan-Einot-Gabriel-Welsch range test (REGWQ) and the Ryan-Einot-Gabriel-Welsch *F* test (REGWF). Both tests utilize a modified significance level based on the number of steps between means computed as:

$$\gamma_r = \begin{cases} 1 - (1 - \epsilon)^{r/k} & \text{if } r < k - 1 \\ \epsilon & \text{if } r \ge k - 1 \end{cases}. \tag{13}$$

The simpler REGWQ test is based on a Studentized range statistic, and a comparison is deemed statistically significant if:

$$\max_{i,j \in R}\left\{\left(\overline{x_i} - \overline{x_j}\right)\sqrt{\min(n_i, n_{ij})}\right\}/s_{pp} \ge S_{\gamma_r, r, f}. \tag{14}$$

The REGWF test is based on an *F*-statistic and is computed as:

[6] For information on how Dunnett's one-tailed *t*-tests are computed, see IBM (2014).

$$\frac{\left(\sum_{i\in R} n_i\overline{x}_i^2-(\sum_{i\in R} n_i\overline{x_i})^2/\sum_{i\in R} n_i\right)}{(r-1)s_{pp}^2} \geq F_{\gamma_r,r-1,f}, \tag{15}$$

where $r = j - i + 1$ and summations are over $R = \{i,\ldots,j\}$. Both the REGWQ and REGWF tests produce homogeneous subset output only.

When assumptions are met, the REGW tests tend to be conservative (Seco, de la Fuente, & Escudera, 2001). Ramsey (1981) found that the REGWQ tended to control Type I error fairly well in an ideal situation where pairs of means were equal and equally spaced from other pairs of means. In other mean configurations, the REGWQ exhibited more power than the Tukey HSD. Unequal sample sizes decreased power and increased Type I errors for the REGWQ, while heterogeneous variances primarily affected Type I error (Demirhan et al., 2009).

PCPs for unequal variances and sample sizes. A smaller set of four tests available in SPSS 23 and 24 do not have the same strict OLS assumptions as the previous PCPs. These tests were created for violation of these assumptions, and thus are theorized to perform adequately in those scenarios. To accommodate heterogeneous variances and sample sizes, variances are weighted by sample size and an estimate is used for the mean square error degrees of freedom. The adjusted degrees of freedom term from Welch (1938) is

$$v = \frac{\left(\frac{s_i^2}{n_i}+\frac{s_j^2}{n_j}\right)^2}{\frac{s_i^4}{n_i^2 v_i}+\frac{s_j^4}{n_j^2 v_j}} \tag{16}$$

where s_i^2 and s_j^2 are the variances for groups *i* and *j*, respectively, and v_i and v_j are the degrees of freedom for groups *i* and *j*, respectively. The weighted variance term only uses the sample sizes and variances from the two groups being compared, and is equal to

$$Q_{i,j}^{*} = \sqrt{\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}}. \tag{17}$$

Tamhane (1979) noted that (at least for the Games-Howell and T2) the tests designed for violation of assumptions may not be as powerful when assumptions are met. However, he notes that the loss of effectiveness is not large when sample sizes are relatively equal. Because the following tests are designed for assumption violations, little research has been conducted on their performance when assumptions are met.

Games-Howell. The Games-Howell (1976) pairwise comparison test is a simple modification of the Tukey HSD that incorporates the adjustments to pooled standard deviation and degrees of freedom. A comparison is statistically significant if the following is true:

$$\left|\overline{x_i} - \overline{x_j}\ \right| \geq Q_{i,j}^{*} S_{\varepsilon,k,v}/\sqrt{2}. \tag{18}$$

When assumptions are met, the Games-Howell procedure controlled Type I error rate fairly well, generally being near the nominal value (Dunnet, 1980b). When assumptions were violated, the Games-Howell procedure was found to be slightly liberal when group sizes were small (i.e., less than 14; Dunnett, 1980b; Tamhane, 1979). Additionally, Demirhan et al. (2009) indicated that heterogeneous variances and unequal group sizes affected comparisons employing the Games-Howell method, although they did not specifically say in what way and indicated that increasing the number of groups seemed to combat these effects. In contrast, Hsiung and Olejnik (1991) found the Games-Howell procedure to adequately control Type I error for data conditions with a 16:1 ratio of largest to smallest variance and a 2:1 ratio of largest to smallest group size. Keselman and Rogan (1978) found similar results with more discrepant variances and group sizes (roughly 22:1 and 3:1, respectively). Further, they concluded that the Games-Howell procedure provided the most power (when compared to several other PCPs) when assumptions were violated.

Tamhane's T2. Tamhane (1979) compared several PCPs designed to control Type I error when variances were heterogeneous. Among those tested were two of his own design: the T1 and T2 procedures. However, "it was demonstrated that T1 is highly conservative relative to T2" (Tamhane, 1979, p. 473), and, consequently, the T2 was deemed the better test and studied further. The T2 is a modified *t*-test deemed significant if:

$$\left|\overline{x_i} - \overline{x_j}\,\right| \geq Q_{i,j}^{*} t_{\gamma,v}, \text{ where } \gamma = 1 - (1 - \epsilon)^{\frac{1}{k^*}}. \tag{19}$$

The modified significance level is based on the Sidak (1967) test, which uses the same adjustment to significance level. Tamhane (1979) suggested a modified version of the T2 (called the T2′ in his paper) for use in certain data conditions when the group sizes and/or group variances are only slightly heterogeneous. However, this modification does not appear as an option in SPSS.

Dunnett (1980b) showed that the T2 tended to be too conservative when group sizes and variances were equal. The power of the T2 test is negatively impacted by unequal group sizes, though less so for larger numbers of groups (Demirhan et al., 2009; Dunnett, 1980b).

Dunnett's C and T3. Dunnett (1980b) extended the work done by Tamhane (1979) and developed two new PCPs for simulation study: C and T3. Dunnett's C is a three or more group extension of Cochran's (1964) solution to the issue of comparing means for two groups with different variances (referred to as the Behrens-Fisher problem). A comparison with Dunnett’s C is statistically significant if:

$$\left|\overline{x_i} - \overline{x_j}\,\right| \geq Q^*_{i,j}\,\frac{\left({S_{\varepsilon,k,n_i-1}s_i^2}/{n_i}+{S_{\varepsilon,k,n_j-1}s_j^2}/{n_j}\right)}{\sqrt{2}\left({s_i^2}/{n_i}+{s_j^2}/{n_j}\right)}. \tag{20}$$

Dunnett examined the performance of the C statistic under heterogeneous variances and group sizes and found it to perform best in moderate to large sample sizes, whereas it was too conservative at small sample sizes. Further, when group sizes and variances were equal, the C tended to be too conservative. Similarly, Hsiung and Olejnik (1999) concluded that the C statistic was conservative with small samples.

Dunnett’s (1980b) T3 is an extension of Tamhane’s T2 that is based on Sidak’s (1967) uncorrelated *t* inequality instead of the multiplicative inequality the T2 uses. Computed with the Studentized maximum modulus, means differ if the following inequality holds:

$$\left|\overline{x_i} - \overline{x_j}\,\right| \geq Q^*_{i,j} M_{\varepsilon,k^*\!,v}. \tag{21}$$

Dunnett found that the T3 was less conservative than Tamhane’s T2 while still controlling Type I error rate for unequal group sizes and variances, though it was conservative when variances and group sizes were equal. Demirhan et al. (2009) found that heterogeneous variances decreased power for the T3.

Research questions

Table 2 contains a summary of the PCPs with some general comments on Type I error and power when assumptions are met and unmet. The research questions addressed in this paper are: (a) in a fully true null hypothesis scenario, which PCP is best suited to maintaining Type I error control when the assumptions of OLS estimation are met and the assumptions of OLS estimation are not met; (b) in a partly true null hypothesis scenario, which PCP is best suited to maximize power while maintaining Type I error control when assumptions are met and assumptions are not met; and (c) in a fully false null hypothesis scenario, which PCP is best suited for maximizing power when assumptions are met and assumptions are not met?

Chapter 3. Method

Simulation

Conditions. Table 3 shows the simulation conditions under four factors: hypothesis, number of groups, group sizes, and group variances. In short, the number of groups were 3, 5, or 7, the sample sizes were either equal at 60 per group or unequal at a ratio of 1:5, and the variances were either equal to 1 or differed by a ratio of 1:7. Two sets of conditions were created when both group sizes and variances were unequal. For one set, the largest variance was paired with the largest group. In the second set, the opposite occurred: the smallest group had the largest variance. Each combination of number of groups, group sizes, and group variances were replicated for fully true null hypotheses (i.e., groups were simulated with equal means to represent coming from the same population), partly true null hypotheses (i.e., groups were simulated to come from one of two populations and groups from the same population were simulated with the same mean), or fully false null hypotheses (i.e., all groups were simulated from different populations with different means). The group means differed depending on the hypothesis condition (more information in the “Data” section below).

There were a total of 45 data conditions. Hypothesis, number of groups, sample size ratio, and variance ratio were crossed (3 X 3 X 2 X 2) for a total of 36 conditions. Additionally, within the cells where both sample sizes and variances were unequal, there were two configurations of the variance ratio: smallest variance with the largest group or smallest variance with the smallest group, adding another 9 cells in the design. Each of these 45 data conditions was replicated 1,000 times. In a similar study, Demirhan et al. (2010) found no difference between 1,000 and 5,000 replications, which informed the decision to complete only 1,000 replications. Across the 45 conditions, there was a total of 45,000 replications.

The decisions made for the group sizes and variances were mostly based on previous literature, but also partly based on good practice. For many of the PCPs studied, the computing limitations at the time of development restricted the sample sizes tested to quite small numbers (e.g. as low as 6 in Games & Howell, 1976). Some recent evaluations of PCPs also maintain small sample sizes (as low as 4 in Demirhan et al., 2010). In part because literature already exists in the field at these small sample sizes and in part because good practice indicates that sampling error of group sizes so small can be detrimental to precision of results, I chose to keep the minimum group size at 20. For unequal group size conditions, the 1:5 ratio results in a 20-100 sample size range. This ratio tends to be slightly larger than what is used in the literature, but is realistic in educational research, where focal groups may be much smaller than comparison groups (e.g., comparing African American students to White students at a primarily White institution). For the variance, the ratio of 1:7 was consistent with the average in the literature, which ranged anywhere from 1:2 to 1:16.

Data. Data were simulated via SAS 9.4 using the "rannor" command, which pulls a random number from a normal distribution with a mean of 0 and a variance of 1[7]. For the fully true null hypothesis conditions, means were set to 10 by adding 10 to each random number, to avoid negative values. For the partly true null hypothesis conditions, one set of means was fixed at 10 and the other group of means was set to a value equal to 0.6 standard deviations above the fixed means (a medium effect size; Cohen 1992). This

[7] SAS syntax for simulating the data is available upon request.

is the standardized mean difference used in Cohen’s *d*, except that Cohen’s *d* is defined by the pooled within-group variances of only two groups. I instead used the square root of the MSE term from the ANOVA as the measure of shared variance to compute the appropriate means corresponding to a standard deviation difference of 0.6. For the fully false null hypothesis conditions, all means differed. The mean for group 1 was fixed at 10 and the mean for the last group was 0.6 standard deviations larger than the first mean. The remaining group means were equally spaced between the smallest and largest means. Consequently, the standard deviation difference between means depended on the number of groups. For example, the mean difference between adjacent group means was smaller in the 7-group case than in the 3-group case.

Given that the variances and sample sizes for groups differ in some conditions, means also differed for conditions. Table 3 shows the means assigned to each group for each condition. The consistency across conditions comes from the repeated maximum difference of 0.6 standard deviations between the group means. For conditions where variances differed between groups, the data were multiplied by the appropriate square root of the variance (i.e., standard deviation; see Table 3) prior to adding the desired mean value.

The final two columns of Table 3 detail the theoretical range of per-pair power and omnibus *F*-test power for the fully true and partly true null hypothesis conditions. The per-pair power values were computed as the theoretical power of independent samples *t*-tests between all simulated groups within a condition. Thus, I expected the 18 PCPs under study to provide slightly lower per-pair power because they were designed to control Type I error and should consequently be less powerful. The omnibus power reported in Table 3 is the theoretical power of the omnibus ANOVA *F*-test to reject the null hypothesis that all group means are equal. I expected that any-pairs power for the 18 PCPs would closely align with the omnibus power.

Procedure and measures

Once data were simulated, a macro was written in SPSS 23 to open the data, run the one-way ANOVA, and output the relevant PCP data to a text file[8]. Syntax for the SPSS macro is available in Appendix A. Then, the text file was read into SAS 9.4 and analyzed. Familywise Type I error rate (a.k.a. experimentwise Type I error rate) was computed for the fully true null hypothesis and partly true null hypothesis conditions, and false discovery rate (FDR) was computed for the partly true null hypothesis (FDR is equal to 1 in fully true null hypothesis conditions and equal to 0 in fully false null hypothesis conditions). Per-pair power, any-pair power, and all-pairs power were computed for the partially true null hypothesis and fully false null hypothesis conditions. The values from the five measures were each aggregated across all relevant replications to provide an average measure value for every PCP.

[8] The macro also runs in SPSS 24 without modification.

Chapter 4. Results

Results are partitioned into three sections that align with my research questions. First, results pertaining to the fully true null hypothesis conditions (i.e., Type I error) are presented. Then, results for the partly true null hypothesis conditions are given (Type I error, FDR, and power). Because it is of secondary interest, only the range of FDR values is provided. Finally, the results from the fully false null hypothesis conditions are provided (power). Within each of the three sections, any general comments about the performance of the PCPs are given before going into specific results for the PCPs.

Fully true null hypothesis conditions

Recall that in these conditions, groups were simulated to come from the same population. Thus, group population means were equal. Tables 4-6 show the Type I error rates for each of the 18 PCPs in the fully true null hypothesis conditions for 3, 5, and 7 groups, respectively. Bolded cells indicate that a test was either too conservative or too liberal when controlling Type I error[9]. Several tests never or almost never adequately controlled Type I error rate at the nominal level (Duncan's MRT, Fisher's LSD, and the Waller-Duncan test) and are thus removed from further consideration in this section.

Assumptions are met. Even when the assumptions of ANOVA are met (Equal N, Equal SD condition), only a subset of tests adequately control Type I error near the nominal level of .05 (Dunnett's *t*, Games-Howell, REGWF, REGWQ, SNK, Tukey's B, and Tukey's HSD). Most of the remaining tests (Bonferroni, Dunnett C, Dunnett T3, Gabriel, Hochberg, Sidak, and Tamhane) controlled Type I error in the 3-group

[9] The acceptable range of values [.037 to .063] was defined as $\alpha \pm 1.96\sqrt{\frac{\alpha(1-\alpha)}{1000}}$, where α was equal to .05. Using the normal approximation to the binomial distribution, this is the 95% confidence interval for a proportion of .05 with 1,000 samples.

conditions but became too conservative as the number of groups increased. The Scheffé test was always too conservative.

Assumptions are not met.

Equal N, Unequal SD. In the 3-group case, all tests controlled Type I error well[10]. However, as the number of groups increased, only the PCPs designed for unequal variances—Dunnett's C, Dunnett's T3, Games-Howell, and Tamhane—maintained Type I error at the nominal rate. All other tests became too liberal with the exception of the Scheffé test, which was again too conservative.

Unequal N, Equal SD. For 3 groups, most tests controlled Type I error well, except for the REGWQ and Scheffé tests that were too conservative. With 5 groups, a large number of tests became too conservative. Only the Gabriel, Games-Howell, REGWF, SNK, Tukey's B, and Tukey's HSD tests maintained adequate Type I error rates. Then, with 7 groups, the Dunnett's C, Dunnett's *t*, Gabriel, Games-Howell, REGWF, and Tukey's HSD performed well. Most other tests were too conservative, except for the SNK and Tukey's B, which were too liberal.

Unequal N, Unequal SD (large). When the largest group had the largest variance, tests tended towards being too conservative. Only Dunnet's C, Dunnet's T3, and Games-Howell maintained appropriate Type I error rates across all three group sizes. The Tamhane test, which is also designed for assumption violations like the above three tests, and Dunnett's *t* were conservative with 3 groups but performed well for 5 and 7 groups. Tukey's HSD controlled Type I error adequately in the 5-group case, but not in the 3- or 7-group conditions.

[10] Excluding Duncan's MRT, Fisher's LSD, and the Waller-Duncan as indicated previously.

Unequal N, Unequal SD (small). When the smallest group had the largest variance, almost every test was too liberal for 3 and 5 groups. Only Dunnett's C, Dunnett's T3, the Games-Howell, and the Tamhane tests adequately controlled Type I error. With 7 groups, Dunnett's C, Dunnett's T3, and the Tamhane all became too conservative, while the Games-Howell continued to control Type I error. Additionally, the REGWQ and the Scheffé tests had acceptable Type I error rates in the 7-group conditions, though this is likely due to their conservative nature overall rather than appropriate control of Type I error. All other tests were too liberal in the 7-group conditions.

Summary. Most tests did not maintain the nominal Type I error rate (within sampling variability) in multiple conditions. The tests designed for assumption violations (Dunnett's C, Dunnett's T3, Games-Howell, and Tamhane) tended to perform better than other tests across the majority of conditions, including when assumptions were met. Only the Games-Howell test had acceptable Type I error rates across all conditions. In the fourth condition, when the largest group had the largest variance, tests were often too conservative. While not problematic in the fully true null hypothesis conditions, power can be negatively affected by Type I error rates that are too conservative. Conversely, in the condition where the smallest group had the largest variance, Type I error rates were often triple or quadruple the nominal level.

Partly true null hypothesis conditions

Moving forward to the partly true null hypothesis conditions, where groups were simulated to come from one of two populations with different means, necessitates consideration of the effects of Type I error control on power. A test can have high power due solely to not controlling Type I error. Essentially, a test could lead researchers to reject virtually every null hypothesis, resulting in high Type I error rates, but simultaneously detect every significant difference, thus having high power. Because researchers can never know which situation their data fall under (i.e., null or alternative distribution), blindly using a test that provides high power at the expense of increased Type I error rates is poor practice. Similarly, using a test that controls Type I error rate so tightly that power is negatively affected is also poor practice. Thus, when I report power statistics for the partly true and fully false null hypothesis conditions, I only do so for the four tests that maintain control over Type I error in all conditions: Dunnett's C, Dunnett's T3, Games-Howell, and Tamhane. Power statistics for all other tests are available in Appendix B[11]. I will continue to report Type I error rates for all tests.

Recall that the power statistics reported are any-pairs power, all-pairs power, and per-pair power. In this study, any-pairs power for a test was computed as the proportion of replications out of 1,000 that correctly identified at least one statistically significant difference. All-pairs power was computed as the proportion of replications out of 1,000 where every statistically significant difference was correctly identified. Per-pair power was computed as the proportion of replications out of 1,000 that correctly identified a statistically significant difference for each individual comparison. Specifically, I report the lowest per-pair power (i.e., the smallest proportion) and the highest per-pair power (i.e., the largest proportion).

[11] These power values should not be interpreted in isolation: instead, always refer back to a test's Type I error rates. Researchers should ask themselves: is my power false (i.e., coming from inflated Type I error rates) or true (i.e., a function of the test)?

The Type I error rates and FDR of the 18 PCPs in the 3-group, 5-group, and 7-group conditions are shown in Tables 7-9, respectively. Largely, tests appeared to control Type I error rate fairly well, at least at the .05 level. In fact, most tests tended to be more conservative than the .05 level because some comparisons for each test in every condition were simulated not to have true differences. Thus, the number of comparisons that can possibly result in a Type I error was reduced to 1 out of 3 for 3 groups, 4 out of 10 for 5 groups, and 9 out of 21 for 7 groups. But because these tests were designed to control Type I error in a familywise manner, the Type I error rates reported were essentially an aggregate of 1, 4, or 9 per comparison Type I error rates when attempting to control familywise error for 3, 10, or 21 comparisons. As a result, tests that have Type I error rates somewhere in the range of .01-.05 were considered as controlling Type I error well.

Assumptions are met. In the Equal N, Equal SD condition, the Duncan, LSD, and SNK tests were all too liberal for all group sizes. When there were 3 groups, both of the REGW tests (REGWQ and REGWF) were also too liberal. All other tests controlled Type I error adequately with 3 groups. Further, with 5 groups, every test but the Duncan, LSD, and SNK controlled Type I error adequately. However, with 7 groups, the Waller-Duncan test also became too liberal. Some tests (the REGWF, REGWQ, and Tukey's B) controlled Type I error near the .05 level, whereas other tests were conservative (Scheffé). FDR values ranged from .010 to .040 for 3 groups, from .002 to .035 for 5 groups, and from .003 to .041 for 7 groups.

Power levels for the Dunnett's C, Dunnett's T3, Games-Howell, and Tamhane tests are shown in Figures 1-3 for 3, 5, and 7 groups, respectively. When assumptions were met, all four tests provided high any-pairs power that increased as the number of groups increased. Intuitively, this makes sense, because as the number of groups increased, so did the number of comparisons being made. Consequently, it was more likely that at least one of the comparisons accurately rejected the null hypothesis. Similar logic, but in reverse, was behind the decrease in all-pairs power from around .65 for 3 groups to around .25 for 7 groups: the more comparisons possible, the less likely to correctly reject all of the null hypotheses that should be rejected. Additionally, as the number of groups increased, the effect size between adjacent means decreased because the smallest group mean was always 0.6 standard deviations lower than the largest group mean with the rest of the groups interspersed equally in between. Per-pair power also decreased from just under .80 to around .60 when moving from 3 to 7 groups. However, the margin between the lowest per-pair power and highest per-pair power remained similar across changes in group sizes.

Assumptions are not met.

Equal N, Unequal SD. For three groups, all tests controlled Type I error below the .05 level. Many tests were quite conservative in their control (< .01), but the Duncan, Dunnett’s C, Dunnett’s T3, Games-Howell, LSD, REGWF, REGWQ, SNK, and Tamhane had less extreme values (> .01). Tests became more liberal with 5 groups, and the Duncan, LSD, and SNK tests no longer controlled Type I error near the .05 level. The REGWF, REGWQ, Tukey’s B, and Waller-Duncan tests all controlled Type I error near the .05 level, and, in some cases, at rates slightly above .05. Results were largely similar between 5 groups and 7 groups, except that Dunnett’s *t* also became slightly too liberal. FDR values ranged from .001 to .012 for 3 groups, from .003 to .034 for 5 groups, and from .003 to .041 for 7 groups.

All-pairs power in the Equal N, Unequal SD condition was nearly identical to the Equal N, Equal SD condition. Any-pairs power was just under .85 for 3 groups, but maxed out around 1.0 (i.e., 100% of replications) for 5 and 7 groups. Interestingly, unlike in the Equal N, Unequal SD condition, as the number of groups increased, so did the gap between the lowest per-pair power and the highest per-pair power. Further, the highest per-pair power was consistently higher and the lowest per-pair power was consistently lower than the previous condition. Highest per-pair power approached 1.0 as the number of groups increased, while lowest per-pair power decreased from roughly .65 to around .30.

Unequal N, Equal SD. The Duncan and SNK tests did not control Type I error rate well with 3 groups. Further, the LSD, REGWF, and Tukey's B had error rates near or slightly above .05. All other tests performed well with 3 groups. With 5 groups, the Duncan, LSD, REGWF, and SNK did not control Type I error. All other tests maintained Type I error rates below or near .05, with Tukey's B and the Waller-Duncan tests close to .05. With 7 groups, the Duncan, LSD, SNK, Tukey's B, and Waller-Duncan tests did not control Type I error. The REGWF performed adequately, but with slightly inflated Type I error rates. All other tests controlled Type I error, with the Scheffé test as the most conservative (possibly too conservative). FDR values ranged from .011 to .056 for 3 groups, from .004 to .046 for 5 groups, and from .001 to .045 for 7 groups.

Any-pairs power was high regardless of the number of groups (> .90). In contrast, all-pairs power began moderately high (around .50) and decreased quickly as the number of groups increased to 5 (around .10) and to 7 (near 0) due to the decreasing effect size between adjacent means. As in the previous condition, the gap between lowest per-pair power and highest per-pair power increased as the number of groups increased. Highest per-pair power remained fairly consistent at slightly below .90, but lowest per-pair power decreased from around .50 at 3 groups to about .15 with 7 groups.

Unequal N, Unequal SD (large). All tests except those designed for violation of assumptions (i.e., Dunnett's C, Dunnett's T3, Games-Howell, and Tamhane) were quite conservative for 3 groups (< .01). For 5 groups, most tests were still conservative, though the Duncan, Dunnett *t*, REGWF, and SNK also had more typical Type I error rates. Additionally, the LSD was too liberal with 5 groups. With 7 groups, the LSD did not control Type I error. Otherwise, tests performed adequately except the Scheffé and Tukey's B, which were still conservative. FDR values ranged from .001 to .009 for 3 groups, from .000 to .014 for 5 groups, and from .001 to .022 for 7 groups.

All measures of power were highest in this condition, which seems counter-intuitive at first. However, recall that power is only reported for tests that are designed for unequal variances and sample sizes. If the power values for the other tests are examined (Appendix B), power tended to be lower for most other tests in this condition. In any event, any-pairs power for the Dunnett's C, Dunnett's T3, Games-Howell, and Tamhane tests was close to 1 regardless of group size. All-pairs power followed established trends of decreasing as the number of groups increased due to decreasing adjacent groups' effect sizes. However, all-pairs power remained around .50 with 5 groups instead of decreasing quite as substantially, as was found in other conditions. Similarly, with 7 groups, all-pairs power, though low at around .10, was still higher than other conditions. Per-pair power was slightly higher than the assumptions met condition (Equal N, Equal SD), but followed the trend of this condition where the gap between lowest per-pair power and highest per-pair power did not increase substantially as the number of groups increased.

Unequal N, Unequal SD (small). Only the Dunnett's C, Dunnett's T3, Games-Howell, and Tamhane tests controlled Type I error for 3, 5, and 7 groups. The Scheffé test had adequate Type I error rates in the 5 and 7 group conditions, though this was likely due to its inherent conservative tendencies. All remaining tests were too liberal. Some tests exceeded the nominal Type I error rate by about .05, but other tests had Type I error rates nearly 10 times the nominal rate. FDR values ranged from .016 to .160 for 3 groups, from .008 to .099 for 5 groups, and from .005 to .090 for 7 groups.

Although any-pairs and highest per-pair power were comparable with other conditions, all-pairs power was at its lowest in this condition, approaching 0 for some tests with 7 groups. Similarly, lowest per-pair power was also at its minimum, near .05 for 7 groups, and consistently lower than other conditions for 3 and 5 groups. These findings appear to be due to the interaction of sample sizes and group variances. Specifically, in this condition with 7 groups, the comparison of the two smallest groups (1 and 2) had the lowest per-pair power. However, in the Unequal N, Unequal SD (large) condition with 7 groups, the lowest per-pair power was when the largest groups were compared (i.e., groups 6 and 7).

Summary. The Duncan, LSD, SNK, and Waller-Duncan tests did not control Type I error in a large number of conditions. Other tests maintained Type I error in most conditions, but were too conservative in the Unequal N, Unequal SD (large) condition and were too liberal in the Unequal N, Unequal SD (small) condition. Only Dunnett's C, Dunnett's T3, the Games-Howell, and the Tamhane tests adequately controlled Type I error in all conditions, similar to the fully true null hypothesis conditions. Power was reported only for these four tests. Any-pairs power was high for all conditions, but all-pairs power decreased as the number of groups increased (because the average effect size between adjacent means decreased) across all conditions. In general, the Unequal N, Unequal SD (large) condition resulted in the highest power levels, and the Unequal N, Unequal SD (small) condition resulted in the lowest power levels. Per-pair power varied greatly across all conditions.

Fully false null hypothesis conditions

Recall that in these conditions, each group was simulated to come from a different population with a different mean. For the fully false null hypothesis conditions, only power was reported for the Dunnett’s C, Dunnett’s T3, Games-Howell, and Tamhane tests, because Type I errors could not be committed. For all fully false conditions, all-pairs power was at or near 0, with the exception of the 3-group, Unequal N, Unequal SD (large) condition, where all-pairs power was at its maximum around .05. Because all-pairs power was universally low, it is not reported for each condition separately. Similarly, lowest per-pair power was at or near 0 in all 5-group and 7-group conditions, and thus is not reported.

Assumptions are met. In the Equal N, Equal SD condition, any-pairs power remained steady at around .80 for 3, 5, and 7 groups. Similarly, highest per-pair power was near .80 with 3 groups, but dropped to around .65 with 5 groups and to around .58 with 7 groups.

Assumptions are not met.

Equal N, Unequal SD. Any-pairs power maintained around .85 regardless of number of groups. Lowest per-pair power began around .17 with 3 groups but immediately dropped to near 0 for 5 and 7 groups. For highest per-pair power, values were around .80 for 3 groups, dropped to around .70 for 5 groups, and ended at around .58 for 7 groups. These results were largely similar to those found when assumptions were met (i.e., the Equal N, Equal SD condition).

Unequal N, Equal SD. Any-pairs power remained fairly consistent at or just below .60 regardless of number of groups in this condition. However, lowest per-pair power began at around .10 for 3 groups and was essentially 0 for 5 and 7 groups. Highest per-pair power was around .50 for 3 groups, decreased to around .35 for 5 groups, and was about .25-.30 for 7 groups.

Unequal N, Unequal SD (large). As was the case in the partly true null hypothesis conditions, the condition where the largest group had the largest variance resulted in the highest power values. Any-pairs power was at or slightly above .90 for 3, 5, and 7 groups. Lowest per-pair power was slightly above .30 for 3 groups, dropped to .03 for 5 groups, and was essentially 0 for 7 groups. Highest per-pair power was at or above .75 regardless of the number of groups, with the 3-group condition resulting in the largest values.

Unequal N, Unequal SD (small). Again similar to the partly true null hypothesis conditions, the lowest power values were observed in conditions where the smallest group had the largest variance. Any-pairs power was slightly above .40 for 3 groups and increased to around .43-.50 for the 5- and 7-group conditions. Lowest per-pair power was

.06 for 3 groups and decreased to essentially 0 for 5 and 7 groups. Highest per-pair power was at its highest for 3 groups at about .25 and decreased to around or slightly under .20 for 5 and 7 groups.

Summary. The power results for the fully false null hypothesis conditions were similar in pattern to the partly true null hypothesis conditions. In general, the fully false null hypothesis conditions resulted in lower power, particularly all-pairs power, likely due to the increased number of comparisons that should have been found significant. That is, as the number of comparisons that should be rejected increased, the more difficult it was to reject all of the appropriate comparisons. I will revisit the issue of all-pairs power again in the Discussion, Limitations, and Conclusions section.

Chapter 5. Discussion, Limitations, and Conclusions

Type I error

The Type I error results from this study are both disappointing and encouraging. First, the vast majority of tests (Bonferroni, Duncan, Dunnett *t*, Gabriel, Hochberg, LSD, REGWF, REGWQ, Scheffé, Sidak, SNK, Tukey B, Tukey HSD, and Waller-Duncan) did not control Type I error when assumptions were violated. In general, increasing the number of groups in the model resulted in worse Type I error rates, whether in the form of too conservative Type I error rates (e.g., Bonferroni, Scheffé) or in the form of too liberal Type I error rates (e.g., LSD, Duncan). Inflated Type I error rates were as high as 59.7% in the fully true null hypothesis conditions and as high as 48.0% in the partly true null hypothesis conditions.

Type I error performance in the Equal N, Unequal SD condition deserves a closer examination, because it is not uncommon to see recommendations to use the better-known procedures if sample sizes are equal but variances may not be, despite the fact that the assumption of homogeneity of variance is violated (Cohen, 2013, chapter 13; Toothaker, 1993). However, I found that as the number of groups increased, many well-known tests became somewhat liberal (e.g., Bonferroni, REGWQ, Tukey's HSD; Tables 4-6). Thus, what has been considered as a relatively "safe" data condition for using the standard PCPs actually results in inflated Type I error.

Fortunately, the four tests designed for violations of assumptions, Dunnett's T3, Dunnett's C, Games-Howell, and Tamhane, controlled Type I error adequately in all conditions. Thus, the adjustment to the degrees of freedom coupled with a weighted pooled variance term based on only the two groups being compared was able to account for the heterogeneity and unequal group sizes simulated in this study. However, Dunnett’s T3, Dunnett’s C, and the Tamhane procedure were each slightly too conservative in at least one condition; thus, the Games-Howell procedure was the test that controlled Type I error the best. I adopt a “better safe than sorry” mentality for PCPs and recommend that researchers and practitioners utilize one of these four tests. Although certain common tests also controlled Type I error when assumptions were met (e.g., Tukey’s HSD), the data condition is unrealistic in real data research (i.e., groups are often unequal and population variances are almost never equal). As the performance of tests such as Tukey’s HSD was not examined under the minor violations of assumptions that are more likely in real data research, simply choosing a test that will control Type I error rate under even more extreme violations is the most logical decision, *if* power is not negatively affected.

If power is not a consideration and controlling Type I error is the only concern, using a more conservative test such as the Bonferroni or Scheffé may be attractive. However, even these tests still had inflated Type I error rates in the Unequal N, Unequal SD (small) condition. Further, recall that in real data research, it is impossible to know if a Type I error is being committed and what the empirical Type I error rate is. Thus, although a researcher could be reasonably sure that the Scheffé test is maintaining the nominal .05 familywise Type I error rate in most situations, he or she cannot be positive. Instead, he or she should use a test that is known to maintain the nominal Type I error rate, such as the Games-Howell procedure. This argument can (and will be) extended to when statistical power *is* a consideration.

Power

Prior to discussing power results from this study specific to the PCPs, I provide some general comments on how the simulation design affected power. In particular, all-pairs power was at or near 0 for most tests in nearly all of the fully false null hypothesis conditions. However, readers should not take this fact to mean that all 18 PCPs are incapable of identifying every statistically significant difference between groups. Rather, readers might question whether there really were "true" differences to find between groups in the fully false null hypothesis conditions. For example, some group means only differed by 0.10 (Equal N, Equal SD, with 7 groups, Table 3). Although these data were simulated to come from different populations, it may be unreasonable to expect a statistical test to identify such seemingly small differences. Why, then, were data simulated as they were?

First, I wanted to keep the largest mean difference between two groups the same as it was in the partly true null hypothesis conditions, 0.6 standard deviations (recall that the square root of the MSE was used for the pooled standard deviation). Doing so provided some level of consistency between the two sets of conditions. Second, and more importantly, all-pairs power had an upper bound defined by lowest per-pair power[12]. In turn, any-pairs power had a lower bound as defined by lowest per-pair power[13]. Thus, if

[12] All-pairs power could not be higher than lowest per-pair power because lowest per-pair was the smallest proportion of correctly rejected tests amongst all comparisons. For example, if a comparison was only found statistically significant in 50% of replications and every other comparison was rejected in those same 50% of replications (or more), all-pairs power would be 50%.

[13] Recall that any-pairs power was the proportion of replications with at least one statistically significant comparison. Thus, any-pairs power could not be lower than lowest per-pair power, because lowest per-pair power was the smallest proportion of statistically significant comparisons amongst all comparisons. For example, if a comparison was found statistically significant in 50% of replications, at least those 50% of replications had one or more statistically significant comparisons, and any-pairs power would be at least 50%.

the simulation were created with larger mean differences between adjacent groups in the fully false null hypothesis conditions, all-pairs power might be higher, but at the expense of any-pairs power being essentially 1.0 in all scenarios. Similarly, as the mean difference between adjacent groups increases, so too does the per-pair power, until highest per-pair power is essentially 1.0 in all scenarios as well. Third, because of its similarity to familywise Type I error, I was focused more on any-pairs power than all-pairs power.

Although power results were only presented for Dunnett's T3, Dunnett's C, Games-Howell, and Tamhane, a more general discussion of power for all tests is required. Specifically, the Type I error control of the Dunnett's T3, Dunnett's C, Games-Howell, and Tamhane tests is not as beneficial if it comes at the cost of lower power when compared to alternative tests. Excluding those tests that do not control Type I error in the majority of conditions (the Duncan, LSD, SNK, and Waller-Duncan) and Dunnett's *t*[14], comparisons between the four tests that controlled Type I error and all others indicate that power for the four tests is roughly the same as any other test (Appendix B). The price for tight Type I error control that the Dunnett's T3, Dunnett's C, Games-Howell, and Tamhane procedures provide is practically insignificantly lower power rates. The REGWF and REGWQ tests did provide meaningfully higher all-pairs and per-pair power in several conditions while simultaneously maintaining Type I error rates, but the lack of Type I error control in several other conditions for these two tests does not make them attractive options.

An argument could be made for purposefully choosing a PCP that provides high power at the expense of Type I error control in exploratory research, where there is no

[14] Dunnett's *t* is excluded because the number of comparisons is different than all other tests. Thus, how any-, all-, and per-pair power is interpreted differs for this PCP comparative to all others.

well-established theory to help dictate comparisons of interest. Exploratory researchers may be more concerned with any statistically significant findings to help inform later follow-up studies. Essentially, these studies are theory-generating studies instead of theory-confirming studies. PCPs are in some ways exploratory by nature because they examine all pairwise comparisons with no *a priori* hypotheses. In this case, inflated Type I error rates may be less of a concern. However, I would argue that instead of switching to a PCP with known higher power but unknown Type I error control, simply increase the nominal Type I error rate from .05 to something like .10 or .15 for the Games-Howell procedure (or Dunnett's C, Dunnett's T3, or Tamhane). Consequently, the experimental Type I error rate is still known, while simultaneously increasing power to account for the experimental nature of the research.

FDR

Although not the main focus of this study, FDR values were computed and ranges were reported for each partly true null hypothesis condition. Recall that none of the 18 PCPs studied were designed to control FDR. Thus, the sometimes large ranges of FDR values were not particularly surprising. In general, tests that had low Type I error rates tended to have lower FDRs, and vice versa for tests with high Type I error rates. The largest issue I find with the FDR statistic is that there is no generally accepted "good" level of FDR, whereas familywise Type I error rates of .05 are considered standard. Further, the total number of computed comparisons influences FDR because of the way FDR is defined.

For example, consider the Games-Howell procedure in the partly true null hypothesis conditions. The Games-Howell controlled familywise Type I error rate around

.20-.26 regardless of number of groups. However, the FDRs for this test were .019 for 3 groups, .009 for 5 groups, and .006 for 7 groups. Familywise Type I error and FDR are not the same. However, it seems as though the main reason for the decrease in FDR while Type I error rate remains fairly constant was related to the number of comparisons being made. Recall that 3 comparisons were made in the 3-group conditions, 10 comparisons were made in the 5-group conditions, and 21 comparisons were made in the 7-group conditions. For procedures such as the Games-Howell that are designed to control Type I error while maximizing power, the increase in possible comparisons means that fewer Type I errors will be made relative to the total number of comparisons, and that the number of correct statistically significant findings (i.e., power) will be maximized. Thus, the numerator of the FDR statistic will remain low while the denominator grows, resulting in lower FDR values with more groups.

Similarly, I find it difficult to interpret FDR. Strictly speaking, the meaning of a .006 FDR is clear: of the total number of comparisons that were rejected, .6% of them were false rejections (i.e., Type I errors). However, I do not know if this is an acceptable rate of false discovery. It seems quite low, but as stated above, that may simply be a function of the total number of comparisons being made. Further, because the FDR is not a particularly well-known statistic, it is reasonable to expect researchers to treat it as a Type I error statistic, and compare values of FDR to a .05 cutoff. However, doing so may be misleading. For example, consider the Tukey B procedure in the partly true null hypothesis, Unequal N, Unequal SD (small) condition with 7 groups. The FDR for this test was .055, which is reasonable when compared to the typical nominal familywise Type I error rate of .05. However, this test had a Type I error rate of .229, over four times the expected nominal rate. Applying the same cutoff to FDR as is typically used for familywise Type I error is clearly not appropriate.

Before readers dismiss FDR as a useless statistic, again recall that these tests were not designed to control FDR. The FDR statistic is likely more interpretable and intuitive to use with procedures that are designed for its use. See Benjamini and Hochberg (1995) or Keselman et al. (1999) for more information about modified PCPs that control FDR instead of Type I error. However, because of how popular and established the concept of Type I error control is for multiple comparisons, I expect FDR will remain a less well-known statistic.

Limitations

Some limitations of this study have already been discussed: the simulation design that essentially resulted in null levels of all-pairs power in the fully false null hypothesis conditions and the reporting of FDR for PCPs that were not designed to control FDR. As with any simulation study, an obvious limitation to this study was the restricted set of conditions. Real data research does not conform to the conditions tested in this study; rather, the conditions tested were intended to be generally similar to real data situations. That is, the simulation conditions were designed to reflect reasonable deviations in group size and variances between groups with the thought that at least some real data group sizes and variances would fall in the same “ballpark” as the simulation.

Next, the data in this study were simulated from a normal population. Rarely is real data perfectly normally distributed. Additionally, I did not examine the effects of dependencies amongst the data like what may be found with cluster sampling that results in nested data. Some research has shown that even mild levels of dependence can seriously inflate Type I error rates (Demirhan et al., 2009; Seco et al., 2001). I chose not to simulate non-normal data in part to reduce the scope of this study, but primarily because ordinary least squares (OLS) estimation tends to be fairly robust to non-normality. For dependence, ANOVA provides a framework for modeling any cluster effects to combat the nuisance of inflated Type I error[15]. To do so, the effect is simply modeled as a random factor in the ANOVA model. For example, if students are nested within schools, model the school effect by entering a school ID variable as a separate grouping factor. In any event, the recommendation I make for PCP choice should be considered only when data are close to normally distributed and independent.

Conclusions and recommendation

When the assumptions of equal sample sizes and variances were violated, only four tests adequately maintained Type I error rate: Dunnett's C, Dunnett's T3, Games-Howell, and Tamhane. All other tests failed to maintain Type I error rate (Bonferroni, Duncan, Dunnett *t*, Gabriel, Hochberg, LSD, REGWF, REGWQ, Scheffé, SNK, Tukey B, Tukey HSD, Waller-Duncan). To avoid capitalizing on false power via inflated Type I error rates, only the four tests that maintained Type I error rates were further considered for power statistics. All four tests provided similar levels of any-, all-, and per-pair power, with the Games-Howell providing a slight edge. Thus, for strict control of Type I error and acceptable power, I recommend utilizing the Games-Howell procedure with normal and independent data (similar to Keselman & Rogan, 1977). Further research is required for non-normal and dependent data.

[15] The inflated Type I error is actually due to deflated standard errors when the nested nature of data is not accounted for.

Although other tests are attractive due to their higher power, I do not recommend their use, as they do not control Type I error at the nominal level. Instead of choosing a test that provides high power at the expense of unknown empirical Type I error, I suggest instead to utilize the Games-Howell procedure with an increased nominal alpha level. Increasing alpha provides greater power at the expense of more Type I errors, but the Type I error rate will be controlled at the value of the nominal alpha, unlike with tests that do not control Type I error. Using the Games-Howell procedure with a less strict alpha level may be particularly useful for exploratory research, where controlling Type I error at the typical .05 level may be of less importance.

Table 1
Notation and description used for PCPs

Notation	**Description**
k	Number of groups
n_i	Number of observations for group *i*
$\overline{x}_i$	Mean of group *i*
s_i	Standard deviation of group *i*
v_i	Degrees of freedom for group *i*, $n_i - 1$
s_{pp}	Square root of the mean square error
ϵ	Familywise error rate (set at .05 in most cases)
α	Comparison error rate
r	Number of steps between means when ordered
f	Degrees of freedom for mean square error
$v_{i,j}$	Absolute difference between the ith and jth means
k^*	Number of comparisons, $k(k-1)/2$
$Q_{i,j}$	$s_{pp}\sqrt{\frac{1}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$
n_h	Harmonic mean of the sample size $\frac{k}{\sum_{1\le i\le k} n_i{}^{-1}}$
Q_h	$s_{pp}/\sqrt{n_h}$

Note: see IBM (2014) for more details.

Table 2

Summary of PCPs

		Assumptions met		Assumptions not met	
Test	**Original citation(s)**	**Type I error**	**Power**	**Type I error**	**Power**
LSD	Fisher (1935)	Does not control	High	Does not control	-
Bonferroni	Dunn (1959; 1961)	Too conservative, particularly with increased number of comparisons	Low	Unequal group sizes increases error, unequal variances less impactful	Decreased
Sidak	Sidak (1967)	Controlled	Low	-	Decreased
SNK	Student (1927); Newman (1939); Keuls (1952)	Controlled only with 3 groups and too liberal otherwise	Medium	Does not control	Decreased
HSD	Unpublished	Controlled	Medium	Does not control	-
Tukey's B	Unpublished	Controlled only with 3 groups and too liberal otherwise	Medium	Does not control	-
Scheffé	Scheffé (1953)	Too conservative with large number of groups	Low	Increases Type I error, but may still be below nominal	-
MRT	Duncan (1955)	Too liberal, particularly with more than 3 groups	High	Differing group sizes increase Type I error	-
GT2	Hochberg (1974)	Controlled	Low-medium	Can be too conservative or too liberal	-

Table 2

Summary of PCPs - continued

		Assumptions met		Assumptions not met	
Test	**Original citation(s)**	**Type I error**	**Power**	**Type I error**	**Power**
Gabriel	Gabriel (1978)	Tends towards conservative	Low-medium	Generally too conservative unless group sizes differ largely	-
Waller-Duncan	Duncan (1965); Waller & Duncan (1969)	Too conservative when *F* is small and too liberal when *F* is moderate to large	Medium-High	Too liberal	Stays medium-high
Dunnett's *t*-test	Dunnett (1955)	Controlled	Medium	Results are approximate when group sizes differ	-
REGWQ/REGWF	Ryan (1960); Einot & Gabriel (1975); Welsch (1977)	Tends towards conservative	Medium	Increased error	Decreased
Games-Howell*	Games & Howell (1976)	Controlled	High	Can be liberal at very group sizes	Remains high
T2*	Tamhane (1979)	Tends towards conservative	-	Generally controlled	Decreased
C*	Dunnett (1980b)	Tends towards conservative	-	Too conservative in small samples	-
T3*	Dunnett (1980b)	Tends towards conservative	-	Controlled	Decreased

Note: * indicates tests designed for unequal group sizes and variances. The categories of Low, Medium, and High are general assessments of PCP power in relation to other PCPs, and are not meant to be precise descriptors of PCP power.

Table 3

Simulation conditions

H_o	***k***	***n_i***	***μ_i***	***σ_i^2***	**Pairwise power**	**Omnibus power**
Fully true	3	60/60/60	10.00/10.00/10.00	1/1/1	-	-
		60/60/60	10.00/10.00/10.00	1/4/7	-	-
		20/60/100	10.00/10.00/10.00	1/1/1	-	-
		20/60/100	10.00/10.00/10.00	1/4/7	-	-
		20/60/100	10.00/10.00/10.00	7/4/1	-	-
					-	-
	5	60/60/60/60/60	10.00/10.00/10.00/10.00/10.00	1/1/1/1/1	-	-
		60/60/60/60/60	10.00/10.00/10.00/10.00/10.00	1/2.5/4/5.5/7	-	-
		20/40/60/80/100	10.00/10.00/10.00/10.00/10.00	1/1/1/1/1	-	-
		20/40/60/80/100	10.00/10.00/10.00/10.00/10.00	1/2.5/4/5.5/7	-	-
		20/40/60/80/100	10.00/10.00/10.00/10.00/10.00	7/5.5/4/2.5/1	-	-
					-	-
	7	60/60/60/60/60/60/60	10.00/10.00/10.00/10.00/10.00/10.00/10.00	1/1/1/1/1/1/1	-	-
		60/60/60/60/60/60/60	10.00/10.00/10.00/10.00/10.00/10.00/10.00	1/2/3/4/5/6/7	-	-
		20/33/47/60/73/87/100	10.00/10.00/10.00/10.00/10.00/10.00/10.00	1/1/1/1/1/1/1	-	-
		20/33/47/60/73/87/100	10.00/10.00/10.00/10.00/10.00/10.00/10.00	1/2/3/4/5/6/7	-	-
		20/33/47/60/73/87/100	10.00/10.00/10.00/10.00/10.00/10.00/10.00	7/6/5/4/3/2/1	-	-
Partly true	3	60/60/60	10.00/10.00/10.60	1/1/1	.91	.93
		60/60/60	10.00/10.00/11.20	1/4/7	.91	.93
		20/60/100	10.00/10.00/10.60	1/1/1	.68-.96	.95
		20/60/100	10.00/10.00/11.39	1/4/7	.68-.96	.95
		20/60/100	10.00/10.00/10.98	7/4/1	.68-.96	.95
	5	60/60/60/60/60	10.00/10.60/10.00/10.60/10.00	1/1/1/1/1	.91	.99
		60/60/60/60/60	10.00/11.20/10.00/11.20/10.00	1/2.5/4/5.5/7	.91	.99
		20/40/60/80/100	10.00/10.60/10.00/10.60/10.00	1/1/1/1/1	.59-.98	.99

Table 3
Simulation conditions – continued

					Pairwise	**Omnibus**
H$_o$	***k***	n_i	μ_i	σ_i^2	**power**	**power**
Partly true	5	20/40/60/80/100	10.00/11.34/10.00/11.34/10.00	1/2.5/4/5.5/7	.59-.98	.99
		20/40/60/80/100	10.00/11.04/10.00/11.04/10.00	7/5.5/4/2.5/1	.59-.98	.99
	7	60/60/60/60/60/60/60	10.00/10.60/10.00/10.60/10.00/10.60/10.00	1/1/1/1/1/1/1	.91	1.0
		60/60/60/60/60/60/60	10.00/11.20/10.00/11.20/10.00/11.20/10.00	1/2/3/4/5/6/7	.91	1.0
		20/33/47/60/73/87/100	10.00/10.60/10.00/10.60/10.00/10.60/10.00	1/1/1/1/1/1/1	.56-.98	1.0
		20/33/47/60/73/87/100	10.00/11.27/10.00/11.27/10.00/11.27/10.00	1/2/3/4/5/6/7	.56-.98	1.0
		20/33/47/60/73/87/100	10.00/11.13/10.00/11.13/10.00/11.13/10.00	7/6/5/4/3/2/1	.56-.98	1.0
Fully false	3	60/60/60	10.00/10.30/10.60	1/1/1	.36-.91	.84
		60/60/60	10.00/10.60/11.20	1/4/7	.36-.91	.84
		20/60/100	10.00/10.30/10.60	1/1/1	.21-.68	.69
		20/60/100	10.00/10.69/11.39	1/4/7	.21-.68	.69
		20/60/100	10.00/10.49/10.98	7/4/1	.21-.68	.69
	5	60/60/60/60/60	10.00/10.15/10.30/10.45/10.60	1/1/1/1/1	.13-.91	.85
		60/60/60/60/60	10.00/10.30/10.60/10.90/11.20	1/2.5/4/5.5/7	.13-.91	.85
		20/40/60/80/100	10.00/10.15/10.30/10.45/10.60	1/1/1/1/1	.09-.69	.73
		20/40/60/80/100	10.00/10.34/10.67/11.01/11.34	1/2.5/4/5.5/7	.09-.69	.73
		20/40/60/80/100	10.00/10.26/10.52/10.78/11.04	7/5.5/4/2.5/1	.09-.69	.73
	7	60/60/60/60/60/60/60	10.00/10.10/10.20/10.30/10.40/10.50/10.60	1/1/1/1/1/1/1	.09-.91	.88
		60/60/60/60/60/60/60	10.00/10.20/10.40/10.60/10.80/11.00/11.20	1/2/3/4/5/6/7	.09-.91	.88
		20/33/47/60/73/87/100	10.00/10.10/10.20/10.30/10.40/10.50/10.60	1/1/1/1/1/1/1	.06-.69	.79
		20/33/47/60/73/87/100	10.00/10.22/10.44/10.66/10.89/11.11/11.33	1/2/3/4/5/6/7	.06-.69	.79
		20/33/47/60/73/87/100	10.00/10.18/10.35/10.53/10.70/10.88/11.06	7/6/5/4/3/2/1	.06-.69	.79

Note. Power of 1.0 indicates power values larger than .995, thus rounding to 1.0.

Table 4

Fully true null hypothesis type I error rates, by condition and test for 3 groups

	Condition				
Test	**Equal N, Equal SD**	**Equal N, Unequal SD**	**Unequal N, Equal SD**	**Unequal N, Unequal SD (large)**	**Unequal N, Unequal SD (small)**
Bonferroni	.041	.042	.038	**.013**	**.167**
Duncan	**.089**	**.095**	**.109**	**.001**	**.297**
Dunnett C*	.049	.043	.043	.037	.046
Dunnett T3*	.044	.040	.039	.036	.043
Dunnett *t*	.049	.061	.057	**.017**	**.159**
Gabriel	.041	.044	.051	**.014**	**.192**
Games-Howell*	.050	.044	.046	.043	.047
Hochberg	.041	.044	.038	**.013**	**.170**
LSD	**.108**	**.127**	**.124**	**.034**	**.286**
REGWF	.050	.048	.051	**.012**	**.175**
REGWQ	.044	.050	**.016**	**.004**	**.096**
Scheffé	**.034**	.037	**.031**	**.010**	**.157**
Sidak	.041	.043	.038	**.013**	**.167**
SNK	.044	.050	.056	**.000**	**.213**
Tamhane*	.043	.039	.039	**.036**	.043
Tukey B	.044	.050	.056	**.000**	**.213**
Tukey HSD	.043	.050	.044	**.015**	**.178**
Waller-Duncan	.040	.041	**.028**	**.000**	**.162**

Note. (large) indicates that the largest group has the largest variance. (small) indicates that the smallest group has the largest variance. Variances used for the 3-group conditions were 1, 4, and 7. Sample sizes for the 3-group conditions were 60 each for equal sizes and 20, 60, and 100 for unequal sizes. *indicates that SPSS labels this post-hoc test as not assuming equal variances. Tests not starred assume equal variances.

Table 5

Fully true null hypothesis type I error rates, by condition and test for 5 groups

	Condition				
Test	**Equal N, Equal SD**	**Equal N, Unequal SD**	**Unequal N, Equal SD**	**Unequal N, Unequal SD (large)**	**Unequal N, Unequal SD (small)**
Bonferroni	**.035**	**.067**	**.032**	**.031**	**.148**
Duncan	**.169**	**.196**	**.169**	.042	**.403**
Dunnett C*	**.034**	.055	**.034**	.052	.047
Dunnett T3*	**.031**	.045	**.031**	.051	.041
Dunnett *t*	.041	**.093**	**.036**	.044	**.117**
Gabriel	.037	**.067**	.037	**.035**	**.162**
Games-Howell*	.040	.058	.039	.060	.052
Hochberg	.037	**.067**	**.033**	**.031**	**.150**
LSD	**.270**	**.279**	**.242**	**.177**	**.429**
REGWF	.049	**.076**	.044	**.034**	**.172**
REGWQ	.042	**.072**	**.018**	**.022**	**.080**
Scheffé	**.013**	.040	**.015**	**.015**	**.099**
Sidak	.037	**.067**	**.033**	**.031**	**.150**
SNK	.042	**.072**	.057	**.009**	**.242**
Tamhane*	**.031**	.045	**.031**	.051	.040
Tukey B	.042	**.072**	.057	**.009**	**.242**
Tukey HSD	.041	**.072**	.039	.038	**.162**
Waller-Duncan	**.082**	**.111**	**.073**	**.015**	**.254**

Note. (large) indicates that the largest group has the largest variance. (small) indicates that the smallest group has the largest variance. Variances used for the 5-group conditions were 1, 2.5, 4, 5.5, and 7. Sample sizes for the 5-group conditions were 60 each for equal sizes and 20, 40, 60, 80, and 100 for unequal sizes. *indicates that SPSS labels this post-hoc test as not assuming equal variances. Tests not starred assume equal variances.

Table 6

Fully true null hypothesis type I error rates, by condition and test for 7 groups

	Condition				
Test	**Equal N, Equal SD**	**Equal N, Unequal SD**	**Unequal N, Equal SD**	**Unequal N, Unequal SD (large)**	**Unequal N, Unequal SD (small)**
Bonferroni	**.032**	**.072**	**.034**	**.022**	**.135**
Duncan	**.275**	**.301**	**.300**	.046	**.507**
Dunnett C*	.046	.049	.041	.046	**.036**
Dunnett T3*	.039	.040	**.035**	.044	**.033**
Dunnett *t*	.051	**.114**	.050	.039	**.110**
Gabriel	**.033**	**.073**	.041	**.023**	**.150**
Games-Howell*	.056	.050	.046	.053	.041
Hochberg	**.033**	**.073**	**.035**	**.023**	**.137**
LSD	**.441**	**.445**	**.445**	**.286**	**.597**
REGWF	.052	**.069**	.056	**.024**	**.151**
REGWQ	.043	**.083**	**.016**	**.020**	.061
Scheffé	**.004**	**.021**	**.007**	**.003**	.055
Sidak	**.033**	**.073**	**.035**	**.023**	**.137**
SNK	.043	**.083**	**.084**	**.005**	**.277**
Tamhane*	.039	.039	**.035**	.043	**.033**
Tukey B	.043	**.083**	**.084**	**.005**	**.277**
Tukey HSD	.043	**.083**	.042	**.029**	**.161**
Waller-Duncan	**.121**	**.156**	**.133**	**.018**	**.323**

Note. (large) indicates that the largest group has the largest variance. (small) indicates that the smallest group has the largest variance. Variances used for the 7-group conditions were 1, 2, 3, 4, 5, 6, and 7. Sample sizes for the 7-group conditions were 60 each for equal sizes and 20, 33, 47, 60, 73, 87, and 100 for unequal sizes. *indicates that SPSS labels this post-hoc test as not assuming equal variances. Tests not starred assume equal variances.

Table 7

Partly true null hypothesis Type I error rates and false discovery rates (FDR), by condition and test for 3 groups

	Condition									
	Equal N, Equal SD		Equal N, Unequal SD		Unequal N, Equal SD		Unequal N, Unequal SD (large)		Unequal N, Unequal SD (small)	
Test	Type I error	FDR	Type I error	FDR	Type I error	FDR	Type I error	FDR	Type I error	FDR
Bonferroni	.018	.012	.002	.001	.018	.013	.001	.001	.129	.084
Duncan	.070	.039	.014	.008	.088	.054	.003	.002	.269	.156
Dunnett C*	.019	.012	.018	.012	.023	.016	.014	.008	.018	.017
Dunnett T3*	.019	.012	.016	.011	.023	.016	.015	.008	.017	.016
Dunnett *t*	–	–	–	–	–	–	–	–	–	–
Gabriel	.018	.011	.002	.001	.021	.014	.001	.001	.142	.089
Games-Howell*	.022	.014	.018	.012	.026	.018	.017	.009	.021	.019
Hochberg	.018	.011	.002	.001	.019	.013	.001	.001	.131	.085
LSD	.069	.038	.015	.008	.049	.029	.002	.001	.208	.117
REGWF	.070	.039	.014	.008	.050	.029	.002	.001	.211	.119
REGWQ	.070	.040	.014	.008	.019	.014	.001	.001	.137	.091
Scheffé	.017	.011	.002	.001	.015	.011	.001	.001	.124	.083
Sidak	.018	.012	.002	.001	.019	.013	.001	.001	.131	.085
SNK	.070	.040	.014	.008	.086	.056	.003	.002	.264	.160
Tamhane*	.019	.012	.016	.011	.023	.016	.014	.008	.017	.016
Tukey B	.039	.018	.006	.004	.060	.042	.002	.001	.228	.148
Tukey HSD	.022	.014	.002	.001	.020	.014	.001	.001	.139	.089
Waller-Duncan	.016	.010	.002	.001	.025	.021	.001	.001	.157	.123

Note. (large) indicates that the largest group has the largest variance. (small) indicates that the smallest group has the largest variance. Variances used for the 3-group conditions were 1, 4, and 7. Sample sizes for the 3-group conditions were 60 each for equal sizes and 20, 60, and 100 for unequal sizes. *indicates that SPSS labels this post-hoc test as not assuming equal variances. Tests not starred assume equal variances. Type I error and FDR for Dunnett's *t* cannot be computed because Dunnett's *t* compares groups 1 and 2 to group 3, which were simulated to come from different populations. Thus, no Type I error can be committed.

Table 8

Partly true null hypothesis Type I error rates and false discovery rates (FDR), by condition and test for 5 groups

	Condition									
	Equal N, Equal SD		**Equal N, Unequal SD**		**Unequal N, Equal SD**		**Unequal N, Unequal SD (large)**		**Unequal N, Unequal SD (small)**	
Test	**Type I error**	**FDR**	**Type I error**	**FDR**	**Type I error**	**FDR**	**Type I error**	**FDR**	**Type I error**	**FDR**
Bonferroni	.017	.005	.022	.006	.017	.006	.003	.001	.085	.029
Duncan	.138	.029	.140	.030	.170	.046	.023	.005	.327	.099
Dunnett C*	.024	.006	.024	.007	.021	.007	.016	.003	.020	.009
Dunnett T3*	.021	.006	.017	.005	.018	.007	.015	.003	.016	.008
Dunnett *t*	.025	.017	.039	.028	.031	.018	.012	.007	.082	.047
Gabriel	.017	.005	.022	.006	.023	.007	.003	.001	.102	.036
Games-Howell*	.025	.006	.025	.007	.022	.007	.017	.004	.020	.009
Hochberg	.017	.005	.022	.006	.018	.006	.003	.001	.088	.030
LSD	.169	.035	.162	.034	.186	.043	.069	.014	.327	.092
REGWF	.047	.011	.056	.013	.069	.020	.013	.003	.167	.055
REGWQ	.045	.011	.055	.013	.018	.006	.003	.001	.078	.030
Scheffé	.008	.002	.011	.003	.010	.004	.001	.000	.050	.020
Sidak	.017	.005	.022	.006	.018	.006	.003	.001	.088	.030
SNK	.093	.022	.101	.024	.104	.033	.016	.004	.251	.091
Tamhane*	.020	.005	.017	.005	.018	.007	.015	.003	.016	.008
Tukey B	.044	.011	.054	.013	.050	.018	.001	.000	.193	.075
Tukey HSD	.019	.005	.024	.007	.022	.007	.006	.002	.093	.032
Waller-Duncan	.041	.010	.044	.010	.052	.017	.001	.000	.201	.071

Note. (large) indicates that the largest group has the largest variance. (small) indicates that the smallest group has the largest variance. Variances used for the 5-group conditions were 1, 2.5, 4, 5.5, and 7. Sample sizes for the 5-group conditions were 60 each for equal sizes and 20, 40, 60, 80, and 100 for unequal sizes. *indicates that SPSS labels this post-hoc test as not assuming equal variances. Tests not starred assume equal variances.

Table 9

Partly true null hypothesis Type I error rates and false discovery rates (FDR), by condition and test for 7 groups

	Condition									
	Equal N, Equal SD		**Equal N, Unequal SD**		**Unequal N, Equal SD**		**Unequal N, Unequal SD (large)**		**Unequal N, Unequal SD (small)**	
Test	**Type I error**	**FDR**	**Type I error**	**FDR**	**Type I error**	**FDR**	**Type I error**	**FDR**	**Type I error**	**FDR**
Bonferroni	.018	.003	.034	.005	.016	.004	.010	.002	.094	.021
Duncan	.237	.032	.230	.034	.242	.042	.054	.008	.456	.090
Dunnett C*	.028	.004	.020	.003	.023	.005	.030	.004	.023	.006
Dunnett T3*	.023	.003	.018	.003	.021	.005	.027	.004	.021	.005
Dunnett *t*	.022	.011	.067	.035	.029	.012	.023	.011	.080	.032
Gabriel	.019	.003	.034	.005	.020	.005	.010	.002	.110	.025
Games-Howell*	.031	.004	.021	.003	.028	.006	.035	.005	.026	.006
Hochberg	.019	.003	.034	.005	.016	.004	.010	.002	.094	.021
LSD	.313	.041	.288	.041	.295	.045	.174	.022	.480	.090
REGWF	.053	.008	.058	.010	.059	.013	.025	.005	.178	.042
REGWQ	.055	.008	.066	.011	.024	.007	.015	.003	.104	.027
Scheffé	.003	.001	.005	.001	.003	.001	.003	.001	.038	.010
Sidak	.019	.003	.024	.005	.016	.004	.010	.002	.094	.021
SNK	.096	.016	.111	.019	.118	.030	.014	.003	.282	.069
Tamhane*	.023	.003	.018	.003	.021	.005	.027	.004	.019	.005
Tukey B	.058	.009	.067	.011	.088	.021	.002	.001	.229	.055
Tukey HSD	.025	.003	.036	.006	.021	.005	.011	.002	.110	.024
Waller-Duncan	.081	.011	.097	.014	.122	.024	.010	.002	.280	.058

Note. (large) indicates that the largest group has the largest variance. (small) indicates that the smallest group has the largest variance. Variances used for the 7-group conditions were 1, 2, 3, 4, 5, 6, and 7. Sample sizes for the 7-group conditions were 60 each for equal sizes and 20, 33, 47, 60, 73, 87, and 100 for unequal sizes. *indicates that SPSS labels this post-hoc test as not assuming equal variances. Tests not starred assume equal variances.

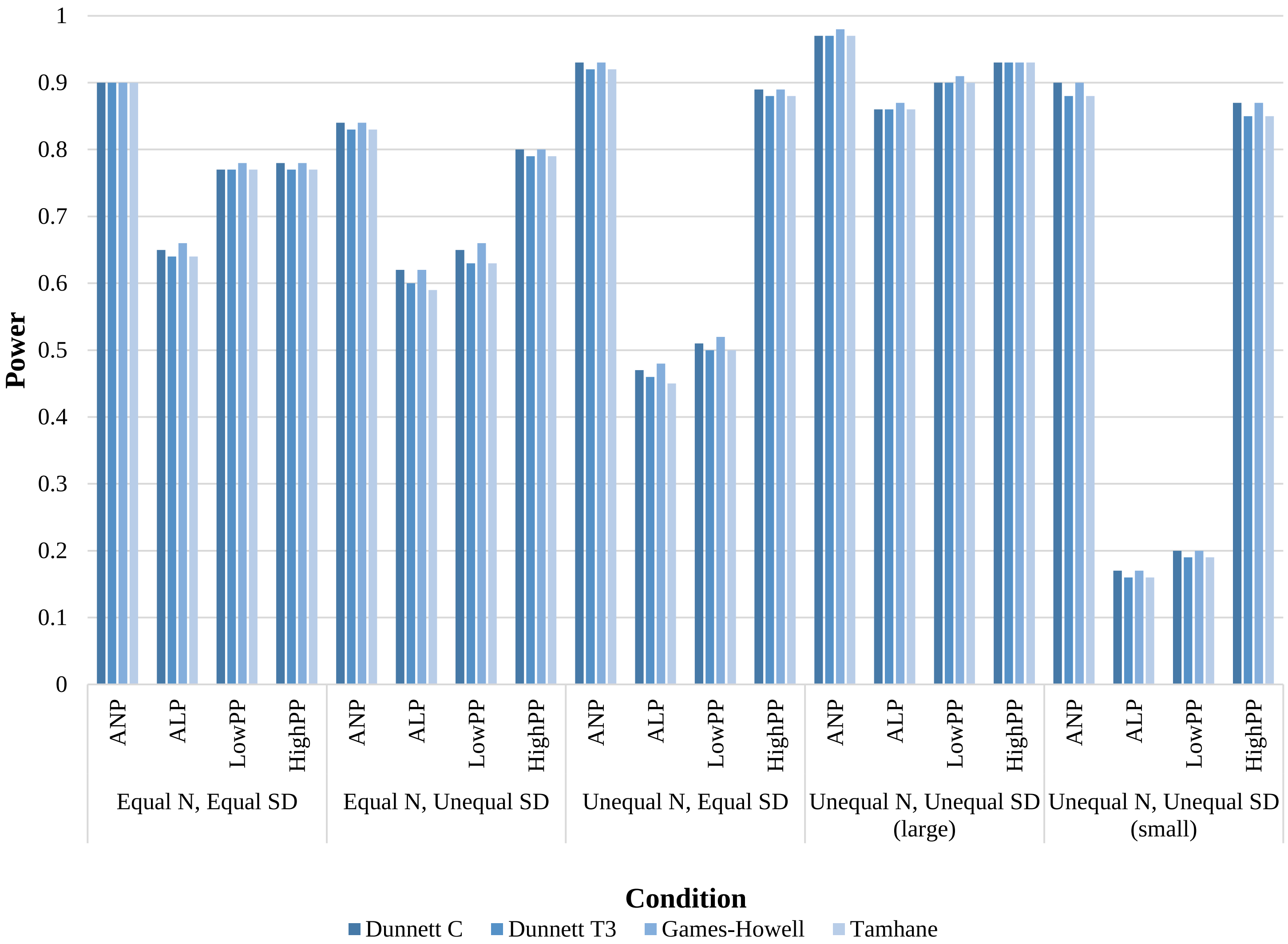

Figure 1. Power values for Dunnett's C, Dunnett's T3, Games-Howell, and Tamhane procedures with 3 groups – partly true null hypothesis conditions. ANP=any-pairs power, ALP=all-pairs power, LowPP=lowest per-pair power, and HighPP=highest per-pair power.

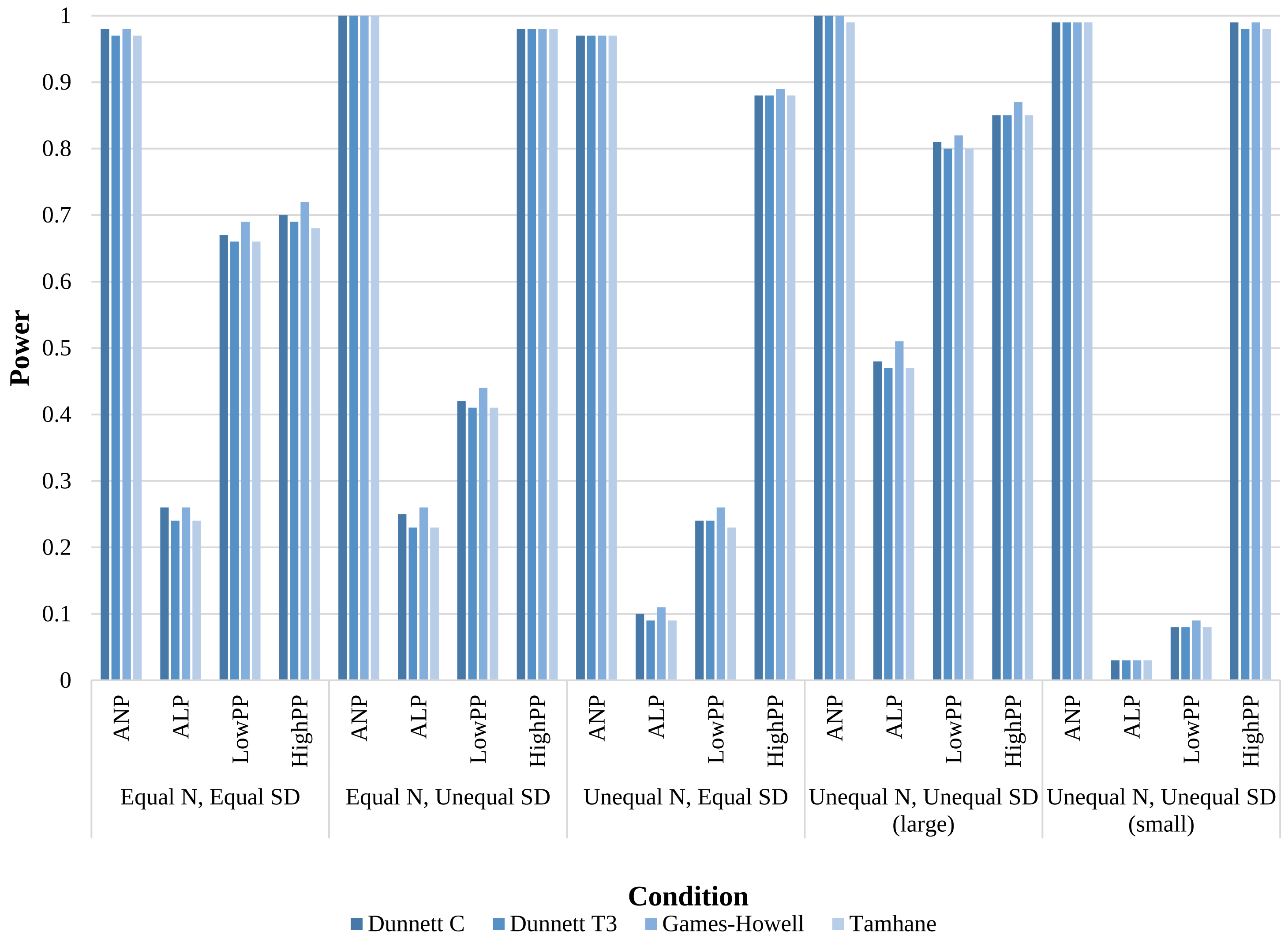

Figure 2. Power values for Dunnett's C, Dunnett's T3, Games-Howell, and Tamhane procedures with 5 groups – partly true null hypothesis conditions. ANP=any-pairs power, ALP=all-pairs power, LowPP=lowest per-pair power, and HighPP=highest per-pair power.

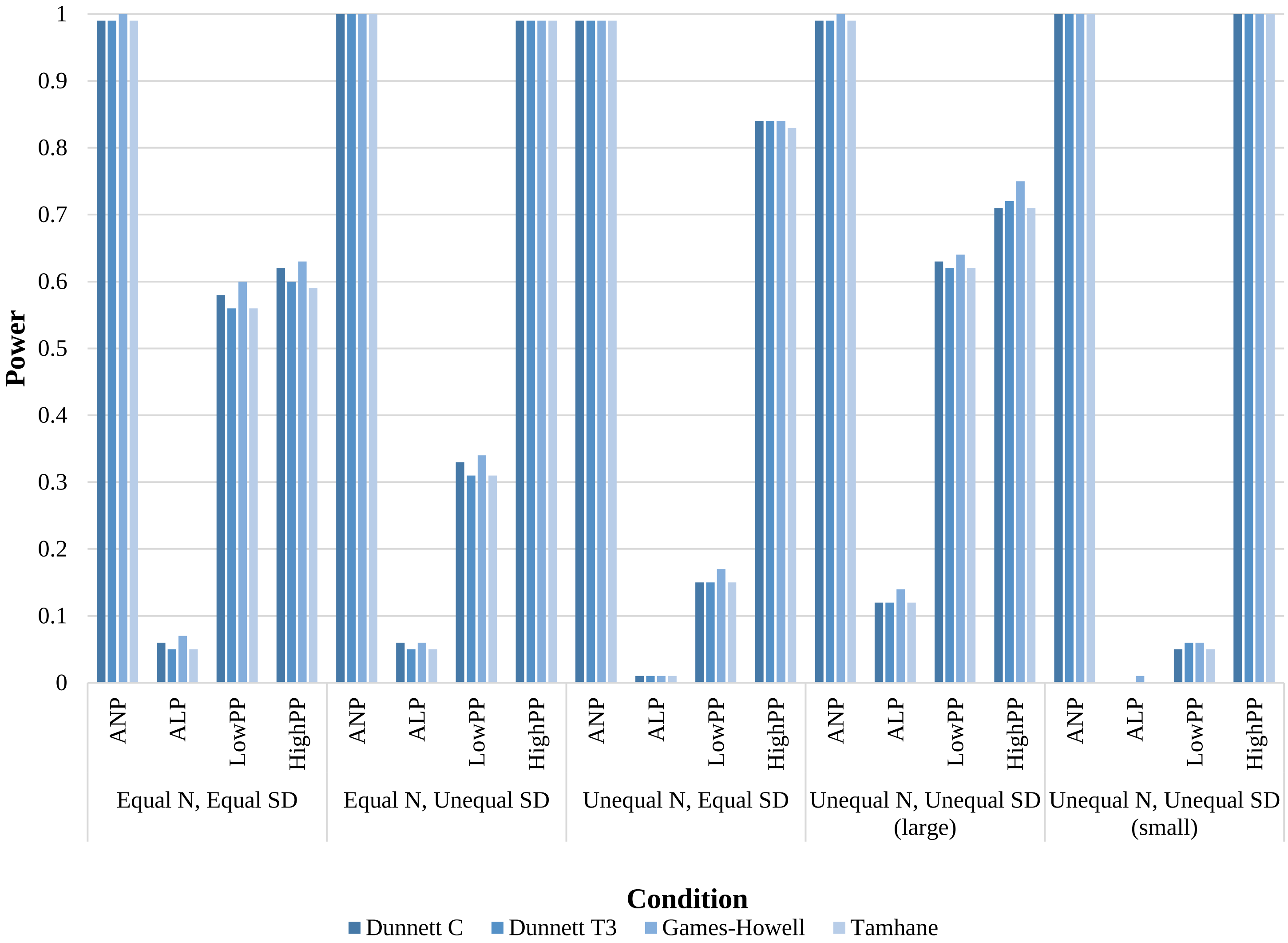

Figure 3. Power values for Dunnett's C, Dunnett's T3, Games-Howell, and Tamhane procedures with 7 groups – partly true null hypothesis conditions. ANP=any-pairs power, ALP=all-pairs power, LowPP=lowest per-pair power, and HighPP=highest per-pair power.

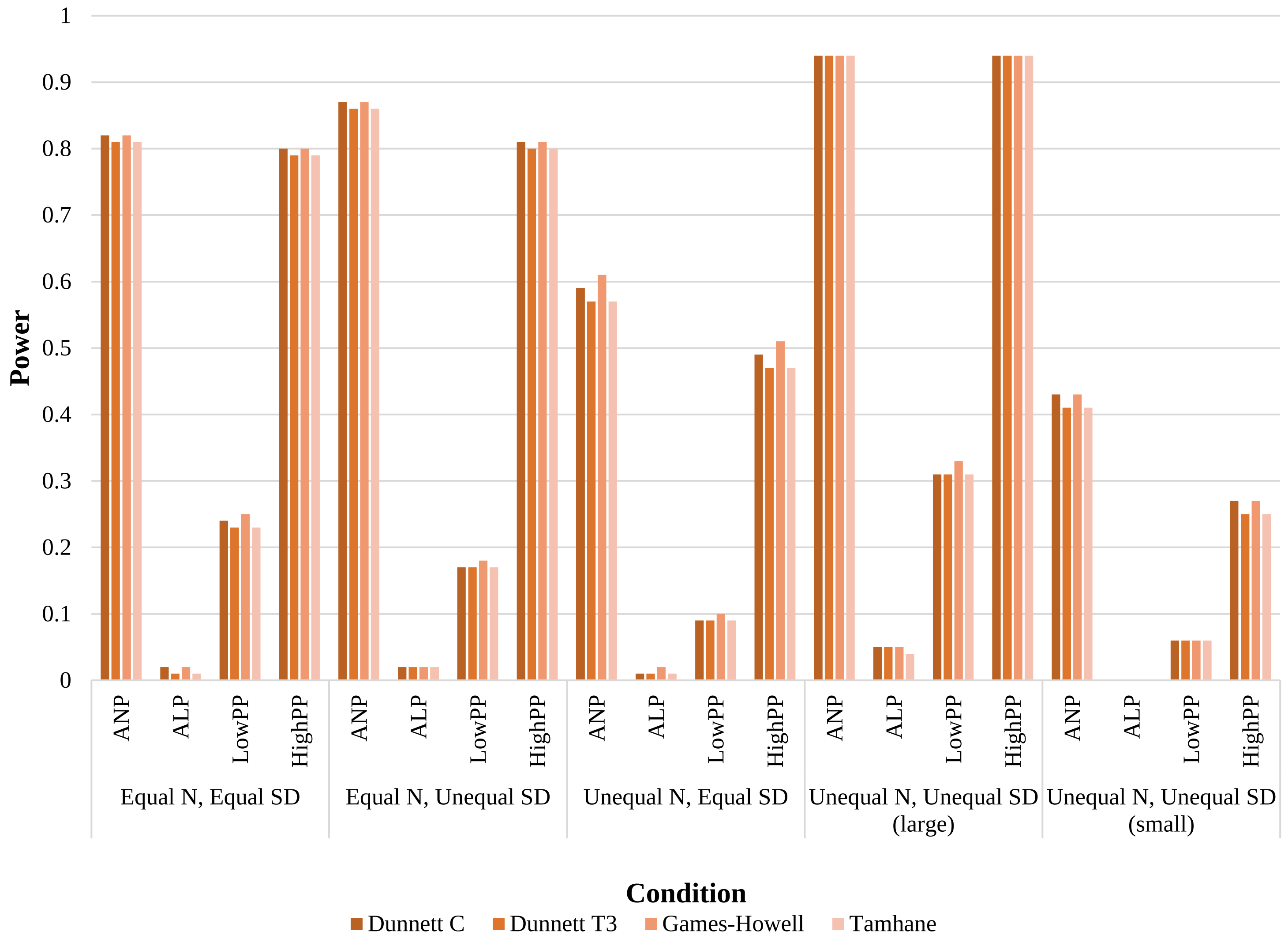

Figure 4. Power values for Dunnett's C, Dunnett's T3, Games-Howell, and Tamhane procedures with 3 groups – fully false null hypothesis conditions. ANP=any-pairs power, ALP=all-pairs power, LowPP=lowest per-pair power, and HighPP=highest per-pair power.

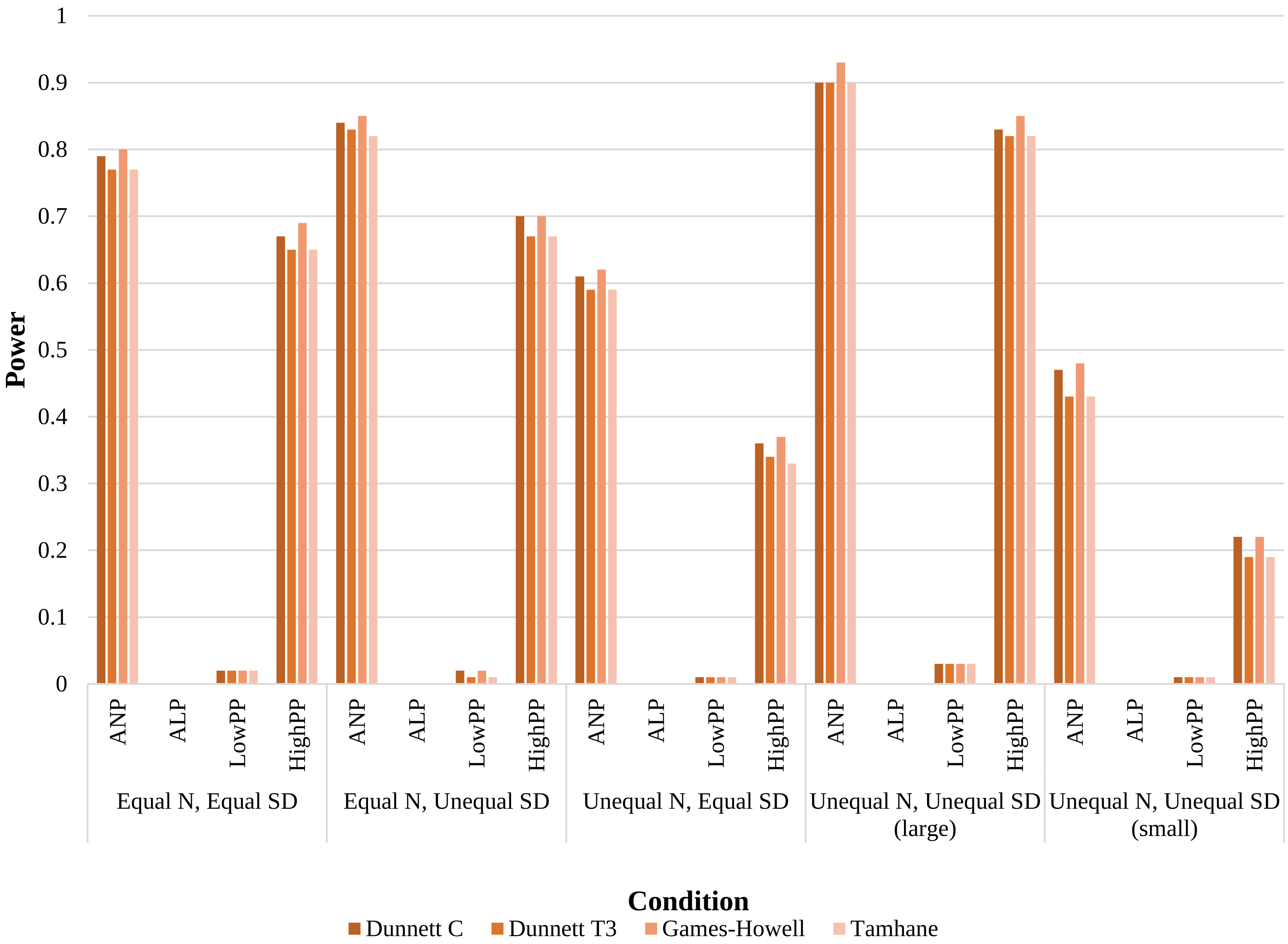

Figure 5. Power values for Dunnett's C, Dunnett's T3, Games-Howell, and Tamhane procedures with 5 groups – fully false null hypothesis conditions. ANP=any-pairs power, ALP=all-pairs power, LowPP=lowest per-pair power, and HighPP=highest per-pair power.

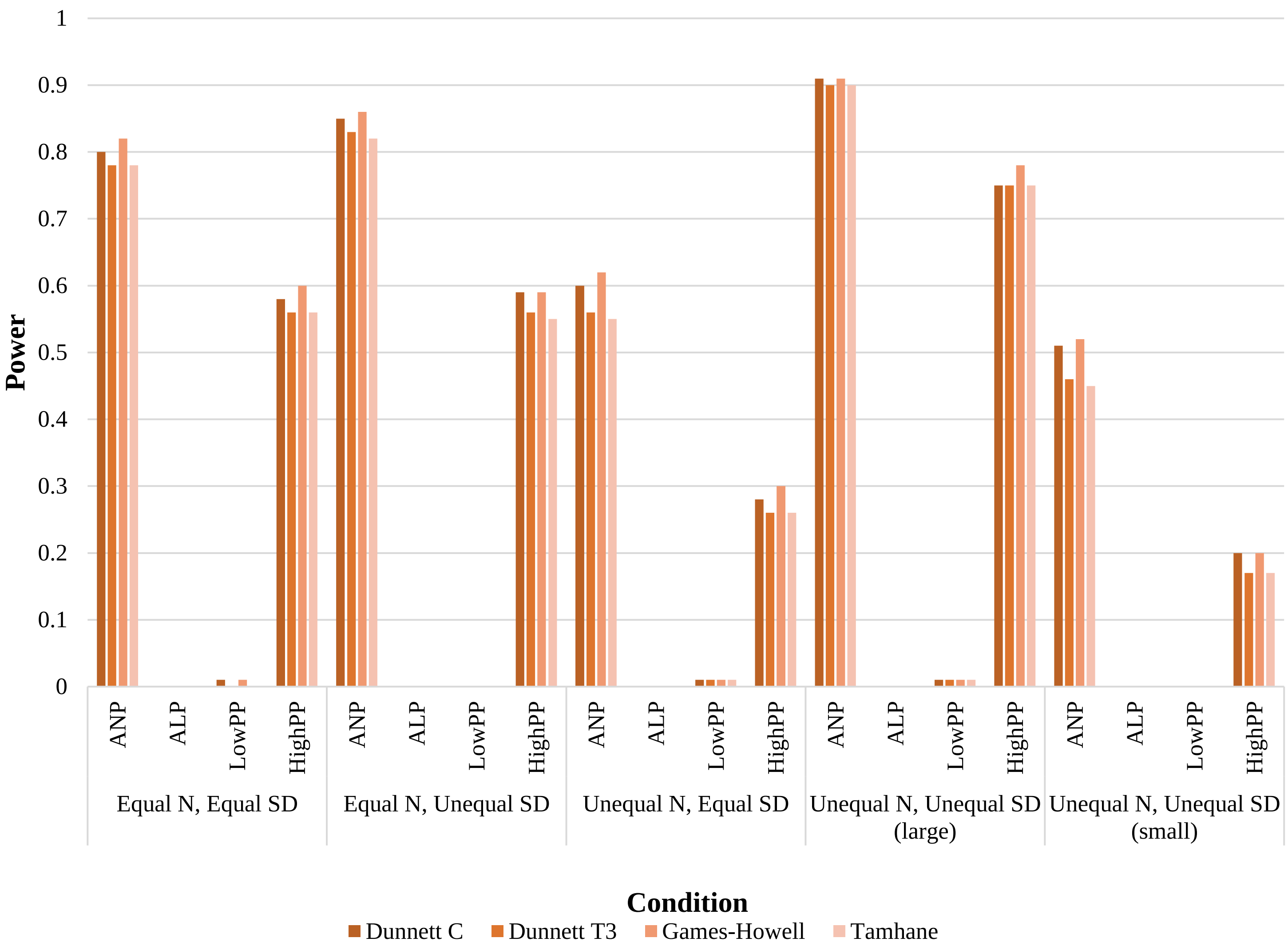

Figure 6. Power values for Dunnett's C, Dunnett's T3, Games-Howell, and Tamhane procedures with 7 groups – fully false null hypothesis conditions. ANP=any-pairs power, ALP=all-pairs power, LowPP=lowest per-pair power, and HighPP=highest per-pair power.

Appendix A

```
*ANOVA macro that opens a data file, computes the ANOVA, and saves the PCP output in a text file.

*This portion of the macro creates the loop and reads in the data.
define !anova (iv=!TOKENS(1) /dv=!TOKENS(1))
!DO !rep = 1 !to 1000.
data list list file= !QUOTE(!CONCAT("FILE PATH"))
 /id grp y.
execute.

*This commend informs SPSS which tables from the output I want sent to a separate text file.
oms
    /select tables
    /destination format = TEXT outfile = !QUOTE(!CONCAT("FILE PATH")) viewer = no
    /if commands = ["oneway"] subtypes = ["multiple comparisons" "homogeneous subsets"].

*Running the one-way ANOVA with all 18 PCPs.
ONEWAY !dv BY !iv
  /MISSING LISTWISE
  /statistics=descriptives
  /POSTHOC=SNK TUKEY BTUKEY DUNCAN SCHEFFÉ LSD BONFERRONI SIDAK GABRIEL
FREGW QREGW GT2 T2 T3 GH C
    WALLER(100) DUNNETT ALPHA(0.05).

*Create separate text files with required output tables and end macro.
omsend.
!DOEND.
!enddefine.

*Running ANOVA macro.
!anova iv=grp dv=y.
```

Appendix B

Table B1

Partly true null hypothesis power rates, by condition and test for 3 groups

	Condition														
	Equal N, Equal SD			Equal N, Unequal SD			Unequal N, Equal SD			Unequal N, Unequal SD (large)			Unequal N, Unequal SD (small)		
Test	**ANP**	**ALP**	**PP**	**ANP**	**ALP**	**PP**	**ANP**	**ALP**	**PP**	**ANP**	**ALP**	**PP**	**ANP**	**ALP**	**PP**
Bonferroni	.89	.65	.77-.78	.86	.70	.76-.80	.92	.49	.54-.88	.91	.50	.52-.89	.94	.47	.52-.89
Duncan	.95	.80	.87-.88	.92	.84	.86-.90	.91	.64	.74-.80	.89	.73	.81	.92	.54	.65-.80
Dunnett C*	.90	.65	.77-.78	.84	.62	.65-.80	.93	.47	.51-.89	.97	.86	.90-.93	.90	.17	.20-.87
Dunnett T3*	.90	.64	.77	.83	.60	.63-.79	.92	.46	.50-.88	.97	.86	.90-.93	.88	.16	.19-.85
Dunnett *t*	.93	.72	.82	.89	.76	.81-.84	.95	.57	.61-.91	.93	.60	.61-.92	.95	.52	.56-.91
Gabriel	.89	.66	.77-.78	.86	.71	.76-.81	.93	.56	.61-.88	.92	.59	.61-.90	.95	.50	.56-.89
Games-Howell*	.90	.66	.78	.84	.62	.66-.80	.93	.48	.52-.89	.98	.87	.91-.93	.90	.17	.20-.87
Hochberg	.89	.66	.77-.78	.86	.71	.76-.81	.93	.49	.54-.88	.91	.50	.52-.89	.94	.47	.52-.89
LSD	.96	.80	.88	.93	.84	.87-.90	.97	.68	.70-.95	.97	.75	.76-.96	.97	.60	.62-.94
REGWF	.91	.79	.85	.88	.83	.84-.87	.95	.70	.72-.93	.95	.78	.79-.94	.95	.61	.63-.93
REGWQ	.90	.78	.84	.87	.81	.83-.85	.88	.51	.53-.86	.84	.53	.53-.84	.89	.48	.50-.87
Scheffé	.88	.63	.75-.76	.84	.68	.75-.78	.91	.47	.52-.86	.90	.46	.48-.88	.93	.45	.50-.88
Sidak	.89	.65	.77-.78	.86	.71	.76-.81	.92	.49	.54-.88	.91	.50	.52-.89	.94	.47	.52-.89
SNK	.90	.78	.84	.87	.81	.83-.85	.83	.62	.70-.75	.80	.69	.74	.86	.52	.63-.76
Tamhane*	.90	.64	.77	.83	.59	.63-.79	.92	.45	.50-.88	.97	.86	.90-.93	.88	.16	.19-.85
Tukey B	.90	.73	.81	.87	.77	.80-.84	.83	.55	.68-.71	.80	.62	.71	.85	.46	.60-.71
Tukey HSD	.90	.67	.78-.79	.87	.72	.77-.82	.93	.52	.56-.89	.92	.53	.55-.90	.95	.49	.54-.90
Waller-Duncan	.89	.634	.76	.85	.68	.75-.79	.78	.37	.57-.58	.73	.43	.57-.58	.81	.31	.55-.57

Note. (large) indicates that the largest group has the largest variance. (small) indicates that the smallest group has the largest variance. Variances used for the 3-group conditions were 1, 4, and 7. Sample sizes for the 3-group conditions were 60 each for equal sizes and 20, 60, and 100 for unequal sizes. *indicates that SPSS labels this post-hoc test as not assuming equal variances. Tests not starred assume equal variances.

Table B2

Partly true null hypothesis power rates, by condition and test for 5 groups

	Condition														
	Equal N, Equal SD			Equal N, Unequal SD			Unequal N, Equal SD			Unequal N, Unequal SD (large)			Unequal N, Unequal SD (small)		
Test	**ANP**	**ALP**	**PP**	**ANP**	**ALP**	**PP**	**ANP**	**ALP**	**PP**	**ANP**	**ALP**	**PP**	**ANP**	**ALP**	**PP**
Bonferroni	.97	.25	.65-.70	.97	.28	.62-.76	.97	.13	.26-.88	.97	.04	.12-.86	.98	.14	.34-.92
Duncan	1.0	.66	.89-.90	1.0	.68	.84-.96	.99	.45	.72-.86	.99	.58	.84-.87	.99	.41	.67-.89
Dunnett C*	.98	.26	.67-70	1.0	.25	.42-.98	.97	.10	.24-.88	1.0	.48	.81-.85	.99	.03	.08-.99
Dunnett T3*	.97	.24	.66-.69	1.0	.23	.41-.98	.97	.09	.24-.88	1.0	.47	.80-.85	.99	.03	.08-.98
Dunnett *t*	.91	.70	.80-.81	.87	.64	.74-.78	.96	.74	.77-.93	.96	.78	.82-.92	.98	.71	.73-.97
Gabriel	.97	.25	.66-.70	.97	.28	.62-.76	.97	.14	.27-.88	.98	.06	.14-.88	.99	.15	.35-.92
Games-Howell*	.98	.26	.69-.72	1.0	.26	.44-.98	.97	.11	.26-.89	1.0	.51	.82-.87	.99	.03	.09-.99
Hochberg	.97	.25	.66-.70	.97	.28	.62-.76	.97	.13	.26-.88	.97	.05	.12-.87	.98	.14	.34-.92
LSD	1.0	.66	.90-.91	1.0	.68	.86-.97	1.0	.42	.58-.98	1.0	.50	.66-.97	1.0	.40	.58-.99
REGWF	.99	.50	.80-.81	1.0	.51	.74-.90	.99	.33	.48-.92	.99	.38	53-.92	.99	.32	.49-.95
REGWQ	.98	.47	.78-.79	.98	.49	.72-.88	.95	.24	.36-.89	.95	.24	.34-.89	.97	.23	.39-.93
Scheffé	.94	.15	.55-.59	.94	.16	.54-.60	.94	.07	.19-.82	.94	.01	.07-.83	.96	.09	.29-.87
Sidak	.97	.25	.66-.70	.97	.28	.62-.76	.97	.13	.26-.88	.97	.05	.12-.87	.98	.14	.34-.92
SNK	.98	.60	.82-.83	.98	.63	.77-.91	.92	.37	.61-.70	.92	.45	69-.71	.92	.35	.60-.73
Tamhane*	.97	.24	.66-.68	1.0	.23	.41-.98	.97	.09	.23-.88	.99	.47	.80-.85	.99	.03	.08-.98
Tukey B	.98	.45	.76-.77	.98	.48	.77-.87	.92	.24	.56-.63	.92	.28	.61.-64	.92	.24	.56-.65
Tukey HSD	.98	.28	.69-.72	.98	.32	.64-.79	.98	.14	.29-.89	.98	.06	.15-.89	.99	.16	.36-.94
Waller-Duncan	.99	.38	.77-.78	1.0	.42	.72-.89	.96	.19	.58-.67	.97	.23	.65-.67	.96	.18	.58-.68

Note. (large) indicates that the largest group has the largest variance. (small) indicates that the smallest group has the largest variance. Variances used for the 5-group conditions were 1, 2.5, 4, 5.5, and 7. Sample sizes for the 5-group conditions were 60 each for equal sizes and 20, 40, 60, 80, and 100 for unequal sizes. *indicates that SPSS labels this post-hoc test as not assuming equal variances. Tests not starred assume equal variances.

Table B3

Partly true null hypothesis power rates, by condition and test for 5 groups

	Condition														
	Equal N, Equal SD			Equal N, Unequal SD			Unequal N, Equal SD			Unequal N, Unequal SD (large)			Unequal N, Unequal SD (small)		
Test	**ANP**	**ALP**	**PP**	**ANP**	**ALP**	**PP**	**ANP**	**ALP**	**PP**	**ANP**	**ALP**	**PP**	**ANP**	**ALP**	**PP**
Bonferroni	.99	.07	.58-.61	.98	.06	.55-.65	.99	.02	.17-.83	.97	.00	.02-.75	1.0	.05	.30-.97
Duncan	1.0	.52	.86-.90	.99	.51	.82-.97	.99	.29	.71-.84	.99	.30	.75-.79	.99	.33	.70-.84
Dunnett C*	.99	.06	.58-.62	1.0	.06	.33-.99	.99	.01	.15-.84	.99	.12	.63-.71	1.0	.00	.05-1.0
Dunnett T3*	.99	.05	.56-.60	1.0	.05	.31-.99	.99	.01	.15-.83	.99	.12	.62-.72	1.0	.00	.06-1.0
Dunnett *t*	.92	.53	.74-.75	.87	.52	.68-.74	.98	.57	.66-.92	.94	.52	.61-.85	1.0	.60	.67-.99
Gabriel	.99	.07	.58-.61	.98	.06	.55-.66	.99	.02	.18-.83	.97	.00	.03-.75	1.0	.06	.31-.97
Games-Howell*	1.0	.07	.60-.63	1.0	.06	.34-.99	.99	.01	.17-.84	1.0	.14	.64-.75	1.0	.01	.06-1.0
Hochberg	.99	.07	.58-.61	.98	.06	.55-.66	.99	.02	.17-.83	.97	.00	.03-.75	1.0	.05	.30-.97
LSD	1.0	.52	.90-.92	1.0	.52	.85-.98	1.0	.28	.56-.98	1.0	.25	.53-.95	1.0	.31	.59-1.0
REGWF	1.0	.24	.73-.76	.99	.25	.65-.85	1.0	.17	.43-.91	.99	.16	.42-.83	.99	.21	.43-.96
REGWQ	.99	.19	.68-.72	.99	.22	.63-.80	.98	.08	.30-.86	.94	.09	.26-.77	.99	.14	.36-.95
Scheffé	.93	.01	.38-.41	.92	.01	.32-.41	.94	.00	.09-.69	.87	.00	.00-.60	.99	.01	.18-.88
Sidak	.99	.07	.58-.61	.98	.06	55-.65	.99	.02	.17-.83	.97	.00	.03-.75	1.0	.05	.30-.97
SNK	.99	.40	.76-.79	.98	.40	.69-.86	.95	.19	.56-.62	.91	.16	.50-.54	.97	.25	.59-.79
Tamhane*	.99	.05	.56-.59	1.0	.05	.31-.99	.99	.01	.15-.83	.99	.12	.62-.71	1.0	.00	.05-1.0
Tukey B	.99	.19	.67-.72	.99	.22	.63-.79	.95	.07	.49-.54	.91	.08	.43-.48	.98	.14	.54-.70
Tukey HSD	1.0	.08	.61-.64	.99	.08	.58-.70	.99	.02	.19-.85	.97	.00	.04-.77	1.0	.07	.32-.98
Waller-Duncan	1.0	.19	.75-.78	.99	.21	.70-.87	1.0	.07	.58-.67	.98	.07	.57-.62	.99	.13	.60-.84

Note. (large) indicates that the largest group has the largest variance. (small) indicates that the smallest group has the largest variance. Variances used for the 7-group conditions were 1, 2, 3, 4, 5, 6, and 7. Sample sizes for the 7-group conditions were 60 each for equal sizes and 20, 33, 47, 60, 73, 87, and 100 for unequal sizes. *indicates that SPSS labels this post-hoc test as not assuming equal variances. Tests not starred assume equal variances.

Table B4

Fully false null hypothesis power rates, by condition and test for 3 groups

	Condition														
	Equal N, Equal SD			**Equal N, Unequal SD**			**Unequal N, Equal SD**			**Unequal N, Unequal SD (large)**			**Unequal N, Unequal SD (small)**		
Test	**ANP**	**ALP**	**PP**	**ANP**	**ALP**	**PP**	**ANP**	**ALP**	**PP**	**ANP**	**ALP**	**PP**	**ANP**	**ALP**	**PP**
Bonferroni	.82	.01	.23-.80	.83	.01	.16-.80	.60	.01	.10-.51	.58	.00	.01-.52	.64	.02	.22-.49
Duncan	.89	.07	.37-.87	.90	.06	.32-.89	.73	.02	.16-.71	.82	.01	.14-.81	.68	.04	.19-.62
Dunnett C*	.82	.02	.24-.80	.87	.02	.17-.81	.59	.01	.09-.49	.94	.05	.31-.94	.43	.00	.06-.27
Dunnett T3*	.81	.01	.23-.79	.86	.02	.17-.80	.57	.01	.09-.47	.94	.05	.31-.94	.41	.00	.06-.25
Dunnett *t*	.85	.28	.29-.83	.85	.32	.32-.85	.67	.23	.32-.58	.68	.31	.35-.64	.69	.19	.35-.54
Gabriel	.82	.01	.24-.80	.83	.01	.16-.80	.65	.01	.12-.58	.67	.00	.02-.63	.68	.02	.24-.54
Games-Howell*	.82	.02	.25-.80	.87	.02	.18-.81	.61	.02	.10-.51	.94	.05	..33-.94	.43	.00	.06-.27
Hochberg	.82	.01	.24-.80	.83	.01	.16-.80	.61	.01	.10-.51	.58	.00	.01-.53	.65	.02	.23-.50
LSD	.91	.07	.38-.89	.93	.06	.33-.90	.78	.06	.20-.68	.82	.01	.06-.79	.77	.08	.31-.60
REGWF	.83	.07	.37-.83	.84	.06	.31-.83	.68	.06	.19-.63	.68	.01	.06-.67	.70	.08	.32-.58
REGWQ	.83	.07	.36-.83	.84	.06	.31-.83	.37	.02	.09-.33	.31	.00	.01-.30	.50	.04	.22-.40
Scheffé	.81	.01	.23-.78	.81	.01	.14-.79	.57	.01	.09-.48	.55	.00	.01-.48	.63	.02	.22-.48
Sidak	.82	.01	.24-.80	.83	.01	.16-.80	.61	.01	.10-.51	.58	.00	.01-.53	.65	.02	.23-.50
SNK	.83	.07	.36-.83	.84	.06	.31-.83	.62	.02	.15-.61	.69	.01	.12-.69	.59	.04	.15-.56
Tamhane*	.81	.01	.23-.79	.86	.02	.17-.80	.57	.01	.09-.47	.94	.04	.31-.94	.41	.00	.06-.25
Tukey B	.83	.04	.30-.82	.84	.04	.23-.82	.62	.01	.11-.61	.69	.00	.08-.69	.59	.02	.12-.56
Tukey HSD	.83	.02	.25-.81	.84	.01	.17-.82	.63	.01	.11-.53	.61	.00	.01-.56	.66	.03	.23-.50
Waller-Duncan	.82	.01	.21-.79	.82	.00	.13-.80	.57	.00	.05-.56	.60	.00	.03-.60	.56	.00	.06-.53

Note. (large) indicates that the largest group has the largest variance. (small) indicates that the smallest group has the largest variance. Variances used for the 3-group conditions were 1, 4, and 7. Sample sizes for the 3-group conditions were 60 each for equal sizes and 20, 60, and 100 for unequal sizes. *indicates that SPSS labels this post-hoc test as not assuming equal variances. Tests not starred assume equal variances.

Table B5

Fully false null hypothesis power rates, by condition and test for 5 groups

	Condition														
	Equal N, Equal SD			Equal N, Unequal SD			Unequal N, Equal SD			Unequal N, Unequal SD (large)			Unequal N, Unequal SD (small)		
Test	**ANP**	**ALP**	**PP**	**ANP**	**ALP**	**PP**	**ANP**	**ALP**	**PP**	**ANP**	**ALP**	**PP**	**ANP**	**ALP**	**PP**
Bonferroni	.79	.00	.01-.67	.79	.00	.00-.68	.62	.00	.01-.36	.52	.00	.00-.34	.69	.00	.02-.40
Duncan	.93	.00	.11-.87	.94	.00	.04-.88	.84	.00	.04-.70	.87	.00	.04-.79	.85	.00	.01-.67
Dunnett C*	.79	.00	.02-.67	.84	.00	.02-.70	.61	.00	.01-.36	.90	.00	.03-.83	.47	.00	.01-.22
Dunnett T3*	.77	.00	.02-.65	.83	.00	.01-.67	.59	.00	.01-.34	.90	.00	.03-.82	.43	.00	.01-.19
Dunnett *t*	.83	.02	.05-.78	.82	.06	.09-.80	.73	.03	.07-.48	.66	.05	.09-.47	.80	.00	.04-.50
Gabriel	.79	.00	.02-.67	.79	.00	.00-.68	.67	.00	.01-.43	.58	.00	.00-.37	.73	.00	.02-.46
Games-Howell*	.80	.00	.02-.69	.85	.00	.02-.70	.62	.00	.01-.37	.93	.00	.03-.85	.48	.00	.01-.22
Hochberg	.79	.00	.02-.67	.79	.00	.00-.68	.63	.00	.01-.36	.53	.00	.00-.34	.69	.00	.02-.40
LSD	.96	.00	.12-.90	.97	.00	.04-.91	.92	.00	.08-.68	.92	.00	.01-.75	.93	.00	.10-.66
REGWF	.84	.00	.05-.74	.85	.00	.01-.75	.73	.00	.03-.56	.71	.00	.00-.59	.77	.00	.03-.55
REGWQ	.82	.00	.05-.73	.82	.00	.01-.74	.38	.00	.01-.26	.27	.00	.00-.20	.48	.00	.02-.31
Scheffé	.68	.00	.01-.57	.68	.00	.00-.57	.49	.00	.00-.26	.36	.00	.00-.22	.58	.00	.00-.32
Sidak	.79	.00	.02-.67	.79	.00	.00-.68	.63	.00	.01-.36	.53	.00	.00-.34	.69	.00	.02-.40
SNK	.82	.00	.07-.74	.82	.00	.02-.74	.63	.00	.02-.54	.61	.00	.02-.56	.69	.00	.00-.55
Tamhane*	.77	.00	.02-.65	.82	.00	.01-.67	.59	.00	.01-.33	.90	.00	.03-.82	.43	.00	.01-.19
Tukey B	.82	.00	.04-.72	.82	.00	.01-.72	.63	.00	.00-.53	.61	.00	.01-.54	.69	.00	.00-.54
Tukey HSD	.82	.00	.02-.70	.82	.00	.00-.71	.66	.00	.01-.39	.56	.00	.00-.36	.71	.00	.02-.43
Waller-Duncan	.88	.00	.04-.78	.89	.00	.01-.80	.74	.00	.01-.61	.74	.00	.01-.65	.76	.00	.00-.58

Note. (large) indicates that the largest group has the largest variance. (small) indicates that the smallest group has the largest variance. Variances used for the 5-group conditions were 1, 2.5, 4, 5.5, and 7. Sample sizes for the 5-group conditions were 60 each for equal sizes and 20, 40, 60, 80, and 100 for unequal sizes. *indicates that SPSS labels this post-hoc test as not assuming equal variances. Tests not starred assume equal variances.

Table B6

Fully false null hypothesis power rates, by condition and test for 7 groups

	Condition														
	Equal N, Equal SD			**Equal N, Unequal SD**			**Unequal N, Equal SD**			**Unequal N, Unequal SD (large)**			**Unequal N, Unequal SD (small)**		
Test	**ANP**	**ALP**	**PP**	**ANP**	**ALP**	**PP**	**ANP**	**ALP**	**PP**	**ANP**	**ALP**	**PP**	**ANP**	**ALP**	**PP**
Bonferroni	.80	.00	.00-.57	.78	.00	.00-.59	.60	.00	.00-.27	.49	.00	.00-.25	.69	.00	.00-.32
Duncan	.97	.00	.07-.87	.97	.00	.01-.86	.93	.00	.01-.71	.92	.00	.01-.80	.92	.00	.00-.65
Dunnett C*	.80	.00	.01-.58	.85	.00	.00-.59	.60	.00	.01-.28	.91	.00	.01-.75	.51	.00	.00-.20
Dunnett T3*	.78	.00	.00-.56	.83	.00	.00-.56	.56	.00	.01-.26	.90	.00	.01-.75	.46	.00	.00-.17
Dunnett *t*	.85	.00	.03-.76	.81	.03	.07-.77	.75	.01	.03-.45	.66	.01	.05-.44	.83	.00	.00-.45
Gabriel	.80	.00	.00-.58	.78	.00	.00-.59	.66	.00	.00-.34	.56	.00	.00-.30	.72	.00	.00-.37
Games-Howell*	.82	.00	.01-.60	.86	.00	.00-.59	.62	.00	.01-.30	.91	.00	.01-.78	.52	.00	.00-.20
Hochberg	.80	.00	.00-.58	.78	.00	.00-.59	.61	.00	.00-.29	.50	.00	.00-.25	.69	.00	.00-.32
LSD	.99	.00	.08-.91	.99	.00	.01-.90	.97	.00	.07-.70	.96	.00	.00-.77	.98	.00	.03-.66
REGWF	.89	.00	.02-.67	.87	.00	.00-.69	.80	.00	.02-.53	.74	.00	.00-.55	.80	.00	.00-.49
REGWQ	.83	.00	.01-.64	.82	.00	.00-.65	.34	.00	.01-.18	.26	.00	.00-.17	.47	.00	.00-.24
Scheffé	.58	.00	.00-.39	.54	.00	.00-.39	.35	.00	.00-.14	.22	.00	.00-.09	.47	.00	.00-.21
Sidak	.80	.00	.00-.58	.78	.00	.00-.59	.61	.00	.00-.29	.50	.00	.00-.25	.69	.00	.00-.32
SNK	.83	.00	.03-.66	.82	.00	.00-.66	.66	.00	.00-.48	.61	.00	.00-.48	.70	.00	.00-.48
Tamhane*	.78	.00	.00-.56	.82	.00	.00-.55	.55	.00	.01-.26	.90	.00	.01-.75	.45	.00	.00-.17
Tukey B	.83	.00	.01-.63	.82	.00	.00-.64	.66	.00	.00-.47	.61	.00	.00-.47	.70	.00	.00-.47
Tukey HSD	.83	.00	.01-.60	.82	.00	.00-.62	.66	.00	.01-.32	.54	.00	.00-.29	.73	.00	.00-.35
Waller-Duncan	.93	.00	.02-.78	.93	.00	.00-.78	.84	.00	.00-.61	.81	.00	.00-.66	.84	.00	.00-.57

Note. (large) indicates that the largest group has the largest variance. (small) indicates that the smallest group has the largest variance. Variances used for the 7-group conditions were 1, 2, 3, 4, 5, 6, and 7. Sample sizes for the 7-group conditions were 60 each for equal sizes and 20, 33, 47, 60, 73, 87, and 100 for unequal sizes. *indicates that SPSS labels this post-hoc test as not assuming equal variances. Tests not starred assume equal variances.

References

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist, 63,* 32-50.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57,* 289-300

Bohrnstedt, G. W. & Carter, T. M. (1971). Robustness in regression analysis. *Sociological Methodology, 3,* 118-146.

Boneau, C. A. (1960). The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin, 57*, 49-64.

Carmer, S. G. & Swanson, M. R. (1973). An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. *Journal of the American Statistical Association, 68,* 66-74.

Cochran, W. G. (1964). Approximate significance levels of the Behrens-Fisher test. *Biometrics, 20,* 191-195.

Cohen, B. H. (2013). *Explaining Psychological Statistics* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.

Curran-Everett, D. (2000). Multiple comparisons: Philosophies and illustrations. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology, 279,* R1-R8.

Demirhan, H., Dolgun, N. A., Demirhan, Y. P., & Dolgun, M. Ö. (2010). Performance of some multiple comparison tests under heteroscedasticity and dependency. *Journal of Statistical Computation and Simulation, 80,* 1083-1100.

Duncan, D. B. (1955). Multiple range and multiple F tests. *Biometrics, 11,* 1-42.

Duncan, D. B. (1965). A Bayesian approach to multiple comparisons. *Technometrics, 7,* 171-222.

Dunn, O. J. (1959). Estimation of the medians for dependent variables. *Annals of Mathematical Statistics, 30,* 192-197.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association, 56,* 52-64.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association, 50,* 1096-1121.

Dunnett, C. W. (1980a). Pairwise multiple comparisons in the homogeneous variance, unequal sample size case. *Journal of the American Statistical Association, 75,* 789-795.

Dunnett, C. W. (1980b). Pairwise multiple comparisons in the unequal variance case. *Journal of the American Statistical Association, 75,* 786-800.

Dunnett, C. W. & Tamhane, A. C. (1992). A step-up multiple test procedure. *Journal of the American Statistical Association, 87,* 162-170.

Einot, I. & Gabriel, K. R. (1975). A study of the powers of several methods of multiple comparisons. *Journal of the American Statistical Association, 70*, 574-583.

Elmore, P. B. & Woehkle, P. L. (1998, April). *Twenty years of research methods employed in "American Educational Research Journal," "Educational Researcher," and "Review of Educational Research."* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Field, A. F. (2013). *Discovering Statistics Using IBM SPSS Statistics* (4th ed.). Los Angeles, CA: Sage.

Fisher, R. A. (1935). *The Design of Experiments.* Edinburgh, London: Oliver and Boyd.

Gabriel, K. R. (1978). A simple method of multiple comparisons of means. *Journal of the American Statistical Association, 73,* 724-729.

Games, P. A. (1971). Multiple comparison of means. *American Educational Research Journal, 8,* 531-565.

Games, P. A. & Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal N's and/or variances: A Monte Carlo study. *Journal of Educational Statistics, 1,* 113-125.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42,* 237-288.

Goodwin, L. D. & Goodwin, W. L. (1985). An analysis of statistical techniques used in the *Journal of Educational Psychology,* 1979-1983. *Educational Psychologist, 20,* 13-21.

Hochberg, Y. (1974). Some generalizations of the *T*-method in simultaneous inference. *Journal of Multivariate Analysis, 4,* 224-234.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75,* 800-802.

Hopkins, K. D. & Weeks, D. L. (1990). Tests for normality and measures of skewness and kurtosis: Their place in research reporting. *Educational and Psychological Measurement, 50,* 717-729.

Hsiung, T. & Olejnik, S. (1991, April). *Power of pairwise multiple comparisons in the unequal variance case.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

IBM (2014). *IBM SPSS Statistics 23 Algorithms.* Retrieved from ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/23.0/en/client/Manuals/IBM_SPSS_Statistics_Algorithms.pdf.

Jaccard, J., Becker, M. A., & Wood, G. (1984). Pairwise multiple comparison procedures: A review. *Psychological Bulletin, 96,* 589-596.

Keselman, H. J., Cribbie, R. A., & Holland, B. (1999). The pairwise multiple comparison multiplicity problem: An alternative approach to familywise and comparisonwise Type I error control. *Psychological Methods, 4,* 58-69.

Keselman, H. J., Cribbie, R. A., & Holland, B. (2004). Pairwise multiple comparison test procedures: An update for clinical child and adolescent psychologists. *Journal of Clinical Child & Adolescent Psychology, 33,* 623-645.

Keselman, H. J., Games, P. A., & Rogan, J. C. (1979). An addendum to “A comparison of the modified-Tukey and Scheffé methods of multiple comparisons for pairwise contrasts.” *Journal of the American Statistical Association, 74,* 626-627.

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

Keselman, H. J. & Rogan, J. C. (1977). The Tukey multiple comparison test: 1953–1976. *Psychological Bulletin, 84,* 1050-1056.

Keselman, H. J. & Rogan, J. C. (1978). A comparison of the modified-Tukey and Scheffé methods of multiple comparisons for pairwise contrasts. *Journal of the American Statistical Association, 73,* 47-52.

Keuls, M. (1952). The use of the "Studentized range" in connection with an analysis of variance. *Euphytica, 1,* 112-122.

Klockars, A. J. & Hancock, G. R. (1992). Power of recent multiple comparison procedures as applied to a complete set of planned orthogonal contrasts. *Psychological Bulletin, 111,* 505-510.

Kromrey, J. D. & La Rocca, M. A. (1995). Power and type I error rates of new pairwise multiple comparison procedures under heterogeneous variances. *Journal of Experimental Education, 63,* 343-363.

Muenchen, R. A. (2016, June 8). *The popularity of data analysis software.* Retrieved from http://r4stats.com/articles/popularity/.

Newman, D. (1939). The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. *Biometrika, 31,* 20-30.

Ozkaya, G. & Ercan, I. (2012). Examining multiple comparison procedures according to error rate, power type and false discovery rate. *Journal of Modern Applied Statistical Methods, 11,* 348-360.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.).Orlando, FL: Harcourt Brace.

Petrinovich, L. F. & Hardyck, C. D. (1969). Error rates for multiple comparison methods: Some evidence concerning the frequency of erroneous conclusions. *Psychological Bulletin, 71,* 43-54.

Ramsey P. H. (1978). Power differences between pairwise multiple comparisons. *Journal of the American Statistical Association, 73,* 479-485.

Ramsey, P. H. (1981). Power of univariate pairwise multiple comparison procedures. *Psychological Bulletin, 90,* 352-366.

Ruxton, G. D. & Beauchamp, G. (2008). Time for some a priori thinking about post hoc testing. *Behavior Ecology, 19,* 690-693.

Ryan, T. A. (1959). Multiple comparison in psychological research. *Psychological Bulletin, 56,* 26-47.

Ryan, T. A. (1960). Significance tests for multiple comparison of proportions, variances, and other statistics. *Psychological Bulletin, 57,* 318-328.

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

Seaman, M. A., Levin, J. R., & Serlin, R. C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures. *Psychological Bulletin, 110,* 577-586.

Seco, G. V., de la Fuente, I. A. M., & Escudero, J. R. (2001). Pairwise multiple comparisons under violation of the independence assumption. *Quality and Quantity, 35,* 61-76.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association, 62,* 626-633.

Stoline, M. R. & Ury, H. K. (1979). Tables of the Studentized maximum modulus distribution and an application to multiple comparisons among means. *Technometrics, 21,* 87-93.

Student. (1927). Errors of routine analysis. *Biometrika, 19,* 151-164.

Tamhane, A. C. (1979). A comparison of procedures for multiple comparisons of means with unequal variances. *Journal of the American Statistical Association, 74,* 471-480.

Toothaker, L. E. (1993). *Multiple comparison procedures.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-089. Newbury Park, CA: Sage.

Waller, R. A. & Duncan, D. B. (1969). A Bayes rule for the symmetric multiple comparisons problems. *Journal of the American Statistical Association, 64,* 1484-1503.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika, 39,* 350-362.

Welsch, R. E. (1977). Stepwise multiple comparison procedures. *Journal of the American Statistical Association, 72,* 566-575.

Winer, B. J. (1971). *Statistical principles in experimental designs* (2nd ed.). New York, NY: McGraw-Hill.