A MORE POWERFUL FAMILYWISE ERROR CONTROL PROCEDURE FOR EVALUATING MEAN EQUIVALENCE

HEATHER P. DAVIDSON

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE MASTER OF ARTS

GRADUATE PROGRAM IN PSYCHOLOGY

YORK UNIVERSITY

TORONTO, ONTARIO

May 2017

ABSTRACT

When one wishes to show that there are no meaningful differences between two or more groups, equivalence tests should be used, as a nonsignificant test of mean difference does not provide evidence regarding the equivalence of groups. When conducting all possible post-hoc pairwise comparisons, *C*, Caffo, Lauzon and Rohmel (2013) suggested dividing the alpha level by a correction of $k^2/4$, where k is the number of groups to be compared, however this procedure can be conservative in some situations. This research proposes two modified stepwise procedures, based on this correction of $k^2/4$, for controlling the familywise Type I error rate. Using a Monte Carlo simulation method, we show that, across a variety of conditions, adopting a stepwise procedure increases power, particularity when a configuration of means has greater than $C - k^2/4$ power comparisons, while maintaining the familywise error rate at or below α. Implications for psychological research and directions for future study are discussed.

TABLE OF CONTENTS

LIST OF TABLES

A More Powerful Familywise Error Controlling Procedure for Evaluating Mean Equivalence

By far the most common way of reporting empirical psychological findings is through the calculation of *p*-values for difference-based statistical tests. These tests involve assessing a research hypothesis of a significant relationship, such as a difference between two means, against a null hypothesis of no relationship. However, the results of these tests do not always provide the proper evidence to support researchers' claims. For example, when the research hypothesis of interest is one of equivalence or a lack of association, traditional difference-based null hypothesis significance testing (NHST) cannot provide evidence of a lack of relationship (i.e., NHST cannot be used in support of the null being true). Recall that an absence of evidence for a difference does not mean there is evidence for an absence of an effect (Altman & Bland, 1995). NHST is based on the probability of a test statistic given that the null hypothesis is true, so researchers' null and alternative hypotheses must properly align with the research questions they wish to answer.

If proper hypotheses are not being evaluated several consequences are possible, including implications for statistical power of the test(s). For example, if a researcher is interested in demonstrating a lack of association but uses a traditional difference-based test, a larger sample size will *decrease* the probability of detecting the negligible relationship, because H_0 will be more likely to be rejected. Additionally, it is unlikely that the true effect is zero (as is specified by the null hypothesis of a traditional difference-based hypothesis), but rather that it is too small to be practically significant in their area of research. This means that researchers' goals should not be to demonstrate a zero effect, but rather that an effect is too small to be considered

meaningful in a practical sense. Thus, alternative procedures known as equivalence tests have been developed to properly address these types of research questions.

Equivalence Testing

Equivalence tests were developed in the biopharmaceutical field for researchers wishing to compare the bioavailability of two drugs (Westlake, 1976; Anderson & Hauck, 1983; Schuirmann, 1987). Researchers needed a way to determine whether a new generic drug was similar enough to an existing brand-name version that it could be prescribed in the place of its more expensive counterpart. This is an example of a research hypothesis of equivalence rather than difference. Equivalence testing is a family of procedures with the goal of detecting a lack of association (e.g., mean equivalence, negligible correlation, lack of interaction). Thus, the null and alternate hypotheses for equivalence tests are effectively the opposite of traditional difference-based hypothesis tests: the null hypothesis states that there is some meaningful relationship among the variables of interest, while the alternative hypothesis states that there is no meaningful relationship. Thus, in equivalence testing a Type I error occurs when one erroneously concludes that there is no meaningful relationship between variables, whereas a Type II error occurs when one erroneously concludes that a relationship is too large to be considered inconsequential. In clinical psychology, for example, researchers often wish to determine whether a treatment group is functionally equivalent to a normal population after the administration of a therapy or drug treatment (Cribbie & Arpin-Cribbie, 2009; Kendall, Marrs-Garcia, Nath & Sheldrick, 1999). These research questions lead to a research hypothesis of equivalence, testing against a null hypothesis that the groups are too different to be considered equivalent. Other possible applications in psychology include evaluating whether two or more experimental groups are equivalent at baseline, whether multiple cultural groups perform

equivalently on some standardized test, whether males and females have equivalent attitudes towards a target, or whether there is a lack of association between two theoretically unrelated predictors in a model.

The two one-sided tests (TOST; Schuirmann, 1987) or confidence interval approach to equivalence testing was introduced to psychology researchers by Rogers, Howard and Vessey (1993), and later by Seaman and Serlin (1998). This approach is the simplest form of equivalence testing and can be used, for example, to determine whether the difference between two population means is small enough that they can be considered equivalent. Evaluating whether means are equivalent using this approach involves first determining the smallest meaningful difference in their area of study, often denoted by δ. Any difference equal to or larger than $|\delta|$ indicates that there is a meaningful difference among the groups, whereas any difference smaller than $|\delta|$ indicates that the difference is too small to be considered meaningful. Because the value δ is meant to quantify what is practically meaningful, researchers choose a value that is theoretically relevant to their research question (Rogers et al., 1993). This can take on the form of a standardized effect size (e.g., Cohen's *d*, Pearson's *r*), a percentage mean difference, or a raw score difference on a well known measure (e.g., some predetermined point difference on the Beck Depression Inventory; Beck, Steer, Bal & Ranieri, 1996). Importantly, this decision must be made *a priori*. Once researchers have determined an appropriate value for δ, they conduct two one-sided *t*-tests with null hypotheses:

$$H_{01}: \mu_1 - \mu_2 \geq \delta; \; H_{02}: \mu_1 - \mu_2 \leq -\delta,$$

where $\mu_1 - \mu_2$ is the difference between the population means of the two independent groups, to determine whether this observed difference is both smaller than δ and larger than $-\delta$. If the null hypothesis is rejected for both tests, then the researchers can conclude that the groups are

equivalent. Stated differently, both null hypotheses are also rejected if the 1-α confidence interval about the mean difference falls within the equivalence interval, $-\delta$ to δ. If not, then the researcher does not have evidence to conclude equivalence. Similar to traditional difference-based NHST, the non-rejection of the TOST cannot be taken as evidence supporting the null hypothesis of difference (Rogers et al., 1993).

One-way Tests of Equivalence

Since their introduction to psychological researchers, equivalence tests have been adapted to fit with different kinds of statistical tests commonly conducted in psychology research. One such example includes tests for comparing multiple independent groups. Wellek (2003) proposed a one-way *F* test to compare three or more group means in an equivalence testing framework, which Cribbie, Arpin-Cribbie and Gruman (2009) showed to be more powerful than the common alternative of conducting all pairwise comparisons and concluding that all groups are equivalent if all pairwise means are declared equivalent. One important thing to note is that one-way *F*-tests, in both a difference and an equivalence framework, can at most tell us that our research hypotheses are partly supported. A significant *F*-statistic from a difference-based ANOVA (i.e., with evidence that supports the research hypothesis) tells us that at least two group means differ from one another, but gives no information regarding which means differ. A non-significant *F*-statistic in a one-way equivalence test (i.e., with evidence that does not support the research hypothesis) tells us that not all of the groups are equivalent; this could mean that none of the groups are similar enough to be deemed equivalent, or that some groups are similar enough while others are not. If one obtains a significant result from a difference-based *F*-test, or a non-significant result from an equivalence-based *F*-test, they must then conduct multiple pairwise comparisons across all of the groups. These comparisons can be made with traditional *t*-tests to

follow up a significant difference-based ANOVA, or using the TOST procedure to follow up a non-significant equivalence-based one-way *F* test. It is important to recognize that, theoretically, rejection of the null hypothesis associated with the omnibus equivalence test provides evidence that all groups are equivalent and hence there is no need to conduct follow-up pairwise multiple comparisons. However, questions have been raised regarding the validity of omnibus tests in equivalence testing (e.g., Cribbie, Ragoonan, & Counsell, 2016) as well as difference-based testing (e.g. Games, 1971; Hancock & Klockars, 1996), and hence we have chosen not to focus on the one-way test of equivalence first, and instead we only discuss the pairwise comparisons.

Familywise Error Rate and Pairwise Comparisons

The problem with conducting post-hoc pairwise tests, each with a Type I error rate α, is that the potential to make a Type I error (i.e., the likelihood of rejecting the null hypothesis when it is in fact true) increases as the number of tests increases. Thus, procedures to maintain the familywise error rate (FWER; α_{FW}), or likelihood of making at least one Type I error across a set of tests, at α have been developed. The simplest method of controlling FWER is the Bonferroni correction (Dunn, 1961), in which the nominal α level is adjusted by the total number of comparisons to be made (e.g., $C = \binom{k}{2}$ for pairwise comparisons, where C is the total number of pairwise comparisons to be made, and k is the number of groups to be compared). This type of correction can either be applied by multiplying each p-value by C, or dividing α_{FW} by C for each comparison such that $\alpha_{PT} = \frac{\alpha_{FW}}{C}$. However, this procedure can be overly conservative, increasing the chance of making a Type II error (i.e., the likelihood of failing to reject the null hypothesis when it is in fact false). Thus, stepwise adaptations of this procedure were developed to decrease the Type II error rate while still maintaining Type I error control below α. Holm's step-down Bonferroni procedure (Holm, 1979), in which p-values are arranged in descending

order and the factor by which the nominal α level (α_{PT}) is adjusted decreases after each significant pairwise comparison, and Hochberg's step-up Bonferroni procedure (Hochberg, 1988), in which *p*-values are arranged in ascending order and the factor by which the nominal α level is adjusted increases after each non-significant pairwise comparison, are examples of such procedures. The rationale of these procedures is that they are less conservative by correcting α_{PT} by a smaller factor each step, while maintaining the FWER at or below α.

As with all equivalence tests, pairwise comparisons require different considerations from an equivalence framework than a difference framework. One important difference is that researchers conducting equivalence tests only need to control for potentially problematic Type I errors (Lauzon & Caffo, 2009; Rohmel, 2011). Any two means that are far enough apart that they would be very unlikely to be mistaken for equivalent (i.e., when a Type I error is highly unlikely) are considered non-problematic and are not controlled for. For example, in a difference-based framework if there was no difference between two means in the population and from our data we conclude that there was a statistically significant difference, then we made a Type I error and any mean difference greater than or less than zero is a problem. However, in an equivalence-based framework, a Type I error is made when one concludes that there is no meaningful difference when in fact there is a meaningful difference. If the true difference in the population has a Cohen's *d*, for example, equal to three, then it is highly unlikely that we would ever conclude that the means are equivalent. However, if the true difference in the population has a Cohen's *d* closer to zero, but the difference is still greater than δ, then we have a greater chance of erroneously concluding that the means are equivalent (i.e., making a Type I error). Only controlling for problematic Type I errors provides more power, thus reducing the chance of making a Type II error. Researchers (e.g., Rohmel, 2011) have defined the area from the

equivalence interval up to twice the equivalence interval (i.e., $|\delta \leq \mu_1 - \mu_1 < 2\delta|$) as a region of problematic Type I errors. As they discussed, any difference in a pair of means falling above this interval is large enough that falsely rejecting the null hypothesis would be highly unlikely. Recall that any difference in a pair of means falling outside of the lower bound of this interval is outside of the boundary of the null hypothesis of $[-\delta, \delta]$, and is instead an instance of statistical power.

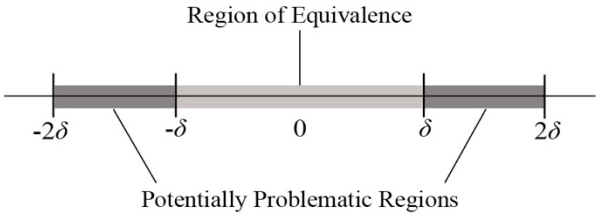

Figure 1: Regions of equivalence and potentially problematic Type I errors

Researchers have attempted to use this region of potentially problematic Type I errors to develop a more powerful FWER control procedure for equivalence tests. Lauzon and Caffo (2009) proposed a Bonferroni-type correction to α_{PT}. According to their proposal, scaling the nominal Type I error rate by a factor of $(k - 1)$, where k is the number of independent groups being evaluated, provides sufficient Type I error control while resulting in a much less conservative rule than a traditional Bonferroni correction of C (i.e., the total number of possible pairwise comparisons). They believe that this factor of $(k - 1)$ only corrects for potentially problematic Type I errors (i.e., those pairs of means with a difference of $|\delta < 2\delta|$), resulting in a more powerful test. The authors noted that the attractiveness of this correction comes with its ease of application, and that it may not be an optimal solution. Rohmel (2011) later showed that while Lauzon and Caffo's correction of $(k - 1)$ works for $k = 3$ (although for reasons Lauzon and Caffo did not consider), it is too liberal for $k \geq 4$. Through a series of proofs, Rohmel showed that a correction of $\alpha/2$ is sufficient to control the FWER for $k = 3$, but that a correction of $\alpha/4$ is

needed to control the FWER for $k = 4$, and a correction of α/6 is needed to control the FWER for $k = 5$.

Building on these two articles, Caffo, Lauzon and Rohmel (2013) proposed a Bonferroni-type correction of $\alpha_{PT} = \alpha_{FW}/(k^2/4)$, where α_{PT} is the per-test alpha level. $k^2/4$ represents the maximum number of comparisons, with k groups, falling in the problematic region of $|\delta \leq \mu_1 - \mu_1 < 2\delta|$. They found that this is a less conservative adjustment than a traditional Bonferroni correction of $\alpha_{PT} = \alpha_{FW}/C$, but still provides adequate FWER control. While this procedure has been shown to control the FWER at approximately α, research on multiple comparisons in a traditional difference-based framework shows that adopting a stepwise approach makes such a test even more powerful while still providing sufficient FWER control (e.g., Holm, 1979; Hochberg, 1998; Keselman, Cribbie & Holland, 2002). This study aims to improve the power of Caffo, Lauzon and Rohmel's correction of $k^2/4$ by adopting a stepwise FWER-controlling procedure.

Current Study

We propose to use a stepwise multiple comparison procedure to provide researchers with a more powerful version of Caffo and colleagues' Bonferroni-type correction (CB), while still controlling the Type I error rate for potentially problematic comparisons. Following the logic of Holm's (1979) step-down procedure or Hochberg's (1988) step-up procedure may provide more power while still maintaining strong control of α_{FW}.

Holm's step-down procedure controls the familywise error rate by adjusting the rejection criteria for each comparison. Let $H_1, \ldots, H_C$ be a family of hypotheses and $P_1, \ldots, P_C$ be their corresponding p-values, with the p-values ordered from smallest to largest. For a given nominal significance level α_{PT}, let j be the minimal value where

$$P_j > \frac{\alpha_{PT}}{C+1-j}. \quad (1)$$

Researchers reject the null hypotheses $H_1, \ldots, H_{(j-1)}$ and fail to reject the null hypotheses $H_j, \ldots, H_C$, where $j = 1, \ldots, C$. If $j = 1$, researchers do not reject any of the null hypotheses. If no value of j satisfies the above equation, then researchers reject all of the null hypotheses. In other words, hypotheses are sequentially compared to a decreasingly adjusted α level until the nominal α level is no longer greater than the observed p-value, at which point no further null hypotheses are rejected.

Recall that $k^2/4$ is the maximum number of potentially problematic comparisons with regards to Type I error (Caffo et al., 2013). In our adjusted Holm procedure (HM), we will test the first $C - k^2/4$ comparisons using a per-test α level of $\alpha_{PT} = \alpha_{FW} / (k^2/4)$. This means that at the test's most conservative level it will only be correcting for the potentially problematic Type I error comparisons that fall within the problematic region of $|\delta \leq \mu_1 - \mu_1 < 2\delta|$. We will then test the remaining $k^2/4$ comparisons using Holm's step-down procedure. This combination of a stepwise procedure with Caffo and colleagues' maximum correction of $k^2/4$ should provide more power than a simple Bonferroni correction of $k^2/4$, while still providing sufficient Type I error control.

Hochberg's step-up procedure follows the same logic but proceeds in the opposite direction. Here, researchers let $H_1, \ldots, H_C$ be a family of hypotheses and $P_1, \ldots, P_C$ be their corresponding p-values, with the p-values ordered from largest to smallest. For a given nominal significance level α, let j be the minimal value where

$$P_j \leq \frac{\alpha}{C+1-j}. \quad (2)$$

Researchers reject the null hypotheses $H_j, \ldots, H_C$ and fail to reject the null hypotheses $H_{(j-1)}, \ldots, H_1$. If $j = 1$, researchers reject all of the null hypotheses. If no value of j satisfies the above

equation, then researchers do not reject any of the null hypotheses. In other words, hypotheses are sequentially compared to an increasingly adjusted α level until the nominal α level is greater than the observed *p*-value, at which point the null hypotheses associated with all remaining (smaller) *p*-values are rejected.

In our adjusted Hochberg procedure (HB), we will test the first $C - k^2/4$ comparisons using Hochberg's step-up procedure, then test the remaining $k^2/4$ comparisons using a per-test α level of $\alpha_{PT} = \alpha_{FW} / (k^2/4)$. In other words, the maximum correction factor in our procedure will again be $k^2/4$, only correcting for comparisons that fall within the problematic region of $|\delta \leq \mu_1 - \mu_1 < 2\delta|$, reflecting the need to only correct for potentially problematic Type I errors. As with HM, HB's combination of a stepwise procedure with Caffo and colleagues' maximum correction of $k^2/4$ will hopefully provide significantly more power than CB while providing sufficient Type I error control. For an example of how these different procedures affect the conclusions of a test (i.e., how each procedure changes the critical α_{PT} for each comparison), see Table 1. Recall that dividing α_{PT} by a given value and comparing the corresponding *p*-value to this corrected α_{PT} is equivalent to multiplying that *p*-value by the same value and comparing the corrected *p*-value to the original α_{PT}. For example, a Bonferroni-type correction can either be applied by multiplying each *p*-value by the correction factor, or dividing α_{PT} by the correction factor for each comparison. For ease of explanation, we frame our example in terms of corrections to α_{PT}.

Simulation Study

This study used Monte Carlo simulations to evaluate the Type I error rates and power of the FWER correction procedures. Using simulations, we compared the CB correction with our proposed HM and HB stepwise procedures, as well as a traditional Bonferroni correction (BF; $\alpha_{PT} = \alpha_{FW} / C$), and no correction for multiplicity (NC; $\alpha_{PT} = \alpha_{FW}$). The δ for all tests was held constant at 20. Although we could have explored alternative values for δ, increasing or decreasing δ has the predictable effect of increasing or decreasing power, respectively. 5000 simulations were conducted for each condition using R version 3.3.1 (R Core Team, 2016), with all pairwise TOSTs being conducted using the equivalencetests package (Cribbie, 2016). A familywise α level of .05 was set for all tests. For each test, FWERs, as well per-pair power rates (the average power across all non-null pairwise comparisons) and all-pairs power rates (the proportion of tests in which all pairs of equivalent means are correctly detected), were computed.

Conditions

We manipulated the number of groups (k), average sample size per group (n), sample size equality/inequality, and population mean configuration. We assessed the effectiveness of these tests using $k = 4$, 7, and 10 independent groups, numbers meant to capture what is typically seen in psychological research. We used average sample sizes of 25 and 50 per cell, representing typical small and moderate per-cell sample sizes in psychology. Group sample sizes were either equal or unequal, with unequal sample sizes either arranged in descending or ascending order. Across conditions, the population within-cell error variance (σ^2) was set at 20. Details regarding the manipulated parameters used in the simulation study are provided in Table 2.

Population mean configurations were chosen to represent various possible combinations of problematic Type I error, non-problematic Type I error and power scenarios. These

configurations include three pure power conditions (i.e., all means falling within the equivalence interval), three conditions with a mix of Type I error and power scenarios, including a "worst case scenario" condition (i.e., the maximum possible problematic Type I error scenarios for the given number of groups), and a pure Type I error condition (i.e., where all means all separated from all other means by $\geq \delta$). See Table 3 for a list of all population mean configurations for each number of independent groups. All mean configurations were crossed with all other variables, resulting in 126 total conditions.

To better understand the worst case scenarios for Type I error rates (i.e., the situation in which the most Type I errors in possible for a given number of groups), let's look at an example. When $k = 7$, $C = 21$ [i.e., $C = \binom{7}{2}$]. One might be tempted to think that the worst case scenario occurs with the greatest number of Type I error scenarios, or in other words, when the difference between all means is $\geq \delta$ (e.g., 0, 20, 40, 60, 80, 100, 120). However, because the difference between many of the pairs of means is $< |2\delta|$), these comparisons are no longer problematic. In this scenario, only 6 out of 21 total pairwise comparisons fall in the region of potentially problematic comparisons. The worst case scenario in terms of Type I errors occurs when the maximum number of comparisons are potentially problematic. This occurs, in this situation, when approximately half of the means are δ greater than the other half (e.g., 0, 0, 0, 0, 20, 20, 20). In this scenario, the differences between 12 out of 21 pairs of means fall in the problematic region of $|\delta \leq \mu_1 - \mu_2 < 2\delta|$, meaning that more than half of the total pairwise comparisons are potentially problematic.

Results

Complete results of the Monte Carlo simulations for all conditions are presented in Tables 4-12 Similar patterns of results (i.e., power and Type I error rates) emerged across levels of equality of sample size (equal or unequal), so they will be discussed together here.

Pure Type I Error Conditions

These mean configurations contained groups that were all separated from each other by $\geq \delta$. FWERs were maintained below $\alpha = .05$ with all three $k^2/4$ correction procedures (CB, HM and HB), with CB, HM and HB producing identical Type I error rates ranging from .01 - .04. In comparison, FWER for the BF ranged from .01 - .03, and for the uncorrected tests ranged from .14 - .28.

Worst Case Scenario Conditions

Type I Error

These mean configurations contained the maximum number of potentially problematic Type I error scenarios possible for the given number of groups, k. With $k = 4$, there is a maximum of 4 problematic Type I error scenarios out of a total of C = 6 pairwise comparisons, when $k = 7$ there is a maximum of 12 problematic Type I error scenarios out of C = 21 pairwise comparisons, and when $k = 10$ there is a maximum of 25 problematic Type I error scenarios out of C = 45 pairwise comparisons. As expected, these configurations produced the greatest FWER, particularly when the number of groups was large, however across all conditions, CB, HM and HB maintained FWER below $\alpha = .05$. This result not only confirms the research of Caffo and colleagues, it also demonstrates that the FWER of the proposed HM and HB do not exceed α. Meanwhile, FWER ranged from .02 - .03 for the BF correction, and from .15 - .50 for the NC comparisons.

Power

In these mean configurations, HM and HB showed consistent but very slight per-pair and any-pairs power advantages over CB (i.e., < 1%). Per-pairs power rates ranged from .38 - .99 for the CB, HM and HB procedures, from .26 - .99 for the BF procedure, and from .93 – 1 for the NC tests. All-pairs power rates ranged from 0 - .99 for the CB, HM and HB procedures, from 0 - .97 for the BF procedure, and from .39 – 1 for the NC tests, with the highest rates seen when k = 4 and n = 50.

Partial Power Conditions

Type I Error

These mean configurations contained some comparisons with mean differences $\geq \delta$ and some comparisons with mean differences $< \delta$. In these configurations, FWER was maintained below α = .05 (between .01 and .04) for the three $k^2/4$ correction procedures (CB, HM, HB). In comparison, Type I error rates for the BF correction ranged from .001 - .02, and for the uncorrected tests ranged from .09 - .29.

Power

Per-pair power rates showed small but consistent increases from the CB correction to the HM and HB corrections. Rates for the HM and HB corrections ranged from .16 – 1, with HB consistently providing slightly more power than HM. As expected, the highest rates were seen when k = 4 and n = 50. Power advantages over the CB correction ranged from 0 – .03 for both the HB and HM procedures, corresponding to 1 - 1.03 times the power. In comparison, per-pair power rates ranged from .11 – .99 with the BF correction, and from .57 – 1 for the NC tests.

All-pairs power rates also showed consistent increases from the CB correction to the HM and HB corrections. Rates for the HM and HB corrections ranged from 0 - .99, with again, as

expected, the highest rates seen when $k = 4$ and $n = 50$. Increases over the CB correction ranged from 0 – .22, corresponding to 1 to 16 times the power for the HM, and 1 to 25 times the power for the HB. In comparison, all-pairs power rates ranged from 0 – .98 with the BF correction, and from 0 – 1 with the NC comparisons.

In these mean configurations, the greatest increases in power, particularly all-pairs power, over the CB procedure were seen when the number of groups was large (i.e., $k = 10$). For example, with means = 0, 0, 0, 0, 0, 0, 0, 0, 0, 20 and $n = 50$, the CB procedure produced an all-pairs power rate of .39, while the HM and HB procedures produced rates of .61. This corresponds to an increase of .22, or 1.57 times the power. In comparison, the BF correction produced an all-pairs power rate of .26, while the NC test produced a rate of .96 (however recall that the FWER for NC was > α).

Pure Power Conditions

These mean configurations consisted of means that all fell within the equivalence interval. The HM and HB corrections showed the greatest advantage over the CB correction when all means fell between 0 and δ. Per-pair power rates showed significant variability, ranging from a low of .21 with the BF correction when $k = 10$, $n = 25$ and the means are further apart (i.e., ranging from 0 – 18), to a high of .98 with the NC tests when $k = 10$, $n = 50$ and the means are closer together (i.e., ranging from 0 – 9). For per-pairs power, the HM and HB corrections consistently produced slightly higher rates than the CB correction, with advantages of up to .12 over CB when $k = 10$, $n = 50$ and the means are closer together.

The greatest power advantage of the HM/HB over the CB were seen for all-pairs power rates. The CB, HM and HB corrections produced maximum all-pairs power rates of .59, .81 and .82, respectively, with the highest rates seen when $k = 4$, $n = 50$ and the means are closer

together. In comparison, maximum all-pairs power rates approached .52 for the BF procedure and .82 for the NC tests under the same conditions. Here, as expected, the HB correction showed identical power rates to NC tests across all conditions (since for the Hochberg to reject all hypotheses the first comparison, using $\alpha_{PT} = \alpha_{FW}$, must be significant which matches the α_{PT} used for the uncorrected procedure), corresponding to power advantages of up to .56 over the CB correction, or in other words, up to 474 times the power of the CB procedure. The HM procedure produced advantages in all-pairs power of up to .46 over the CB procedure, corresponding to about 176 times the power.

For example, with means = 0, 1.5, 3, 4.5, 6, 7.5, 9 and $n = 50$, the CB procedure produced an all-pairs power rate of .20, while the HM procedure produced a rate of .66 and the HB procedure produced a rate of .71. This corresponds to respective increases of .47 and .57, or 3.34 or 3.58 times the power. In comparison, the BF correction produced an all-pairs power rate of .14, while the NC test produced a rate of .71 (however recall that the FWER for NC was > α).

Overall Summary

As expected, across all 126 conditions, NC had the highest power rates, followed by (respectively) the HB correction, the HM correction, the CB correction, and finally BF. Overall, as expected, the highest all-pairs power rates were seen when the number of groups was small (i.e., $k = 4$), the sample size per cell was large (i.e., $n = 50$) and the means were close together (i.e., the mean configuration with the smallest variability for each value of k). The highest per-pair power rates were also seen when the average sample size per cell was large and the means were close together, but there was no consistent pattern with regards to number of groups.

Discussion

Pairwise comparisons from an equivalence testing framework require different considerations than pairwise comparisons from a traditional difference-based framework. Making a Type I error in an equivalence test involves concluding two means are similar enough to be considered equivalent when they are in fact meaningfully different. Although the difference between means can increase to infinity, in practice we only need to control for comparisons in which we have a reasonable chance of making a Type I error. Recall that Rohmel (2011) defined the region of potentially problematic Type I errors as the area from the equivalence interval up to twice the equivalence interval (i.e., $|\delta \leq \mu_1 - \mu_2 < 2\delta|$). Along with Caffo and Lauzon (2013), he showed that by only controlling for differences between means that fall within this region, equivalence tests have the ability to be more powerful than if they controlled for all differences in means, while still maintaining FWER control at or below α.

This study aimed to further increase power by utilizing Holm- and Hochberg-type stepwise correction procedures, while maintaining FWER closer to α. Our results showed that adding a stepwise algorithm increased power over the CB correction while maintaining familywise Type I error rates below α in all configurations. The configurations that showed the most improvement in power fall into two main categories: configurations in which all means are in fact equivalent, and configurations in which some means are equivalent and there are a proportionately large number of equivalent means.

More specifically, our modified stepwise procedures show improved power when the total number of power comparisons (i.e., pairs of means that are in fact equivalent) is greater than ($C - k^2/4$). This is due to the nature of our modifications of the stepwise procedures, which makes the maximum correction $k^2/4$. Instead of stepping down from C to 1 in a Holm-type

procedure or stepping up from 1 to C in a Hochberg-type procedure, our procedure involves stepping down from $k^2/4$ to 1 in the HM procedure and stepping up from 1 to $k^2/4$ in the HB procedure. This means that only the largest ($k^2/4 - 1$) p-values are gaining power over Caffo and colleagues' CB procedure, while the remaining ($C - [k^2/4 - 1]$) are being corrected by the same factor of $k^2/4$. This is different than the traditional Holm or Hochberg correction procedures in a difference-based framework, where all but the smallest p-value are being corrected by a smaller factor than with a Bonferroni correction. By this logic, a configuration of means must have greater than ($C - k^2/4$) power comparisons in order for the HM or HB correction to provide increased power over the CB correction.

In practical terms, this fact manifests itself in two ways. First, if not all means are in fact equivalent, the number of groups, k, must be greater than seven for our HM and HB corrections to provide more power than the CB correction. This is the minimum number of groups with which there can be more than ($C - k^2/4$) power comparisons without all means falling within the equivalence interval. As our results show, the greatest increases in power over the CB procedure when not all means were equivalent were seen with 10 groups. While all-pairs power rates are generally lower with a large number of groups - as k increases, so does C, so it becomes harder to detect equivalence between all pairwise comparisons - these configurations provided the greatest opportunity for the HM and HB stepwise procedures to increase power over the traditional Bonferroni-type CB procedure.

Second, our HM and HB corrections are always more powerful than the CB correction when all means are in fact equivalent. This is because when all means are equivalent, ($k^2/4 - 1$) out of C total comparisons will be adjusted by a more liberal factor with a stepwise correction than with a Bonferroni-type correction such as the CB. With our HB correction, all-pairs power

rates are as high as with no correction at all, while still maintaining Type I error rates below - and close to - α, which is an extremely meaningful benefit to using a stepwise procedure. That being said, if the means are in fact equivalent and one conducts an omnibus equivalence test as a gatekeeper before conducting pairwise comparisons, the omnibus test will likely be rejected and pairwise comparisons will therefore be unnecessary, stopping the analytic process before a Holm- or Hochberg-type procedure has a chance to show its greatest benefit. However, there is research to suggest that using an omnibus test as a gatekeeper is unwise. Cribbie, Arpin-Cribbie and Gruman (2009) concluded that a one-way *F* test (e.g. Wellek, 2003) is recommended over conducting all pairwise comparisons in an equivalence framework because existing approaches to conducting all pairwise comparisons were overly conservative. However, as mentioned earlier, Cribbie, Ragoonanan and Counsell (2016) explain that the omnibus test can sometimes be incoherent with pairwise comparisons, allowing lower-order (e.g., pairwise) differences larger than δ to be declared equivalent. That being said, if all pairwise comparisons are significant, the omnibus test will be too. As Hancock and Klockars (1996) point out, the omnibus test is rarely of substantive interest and serves instead to provide Type I error control, which makes this test redundant if a pairwise comparison procedure exists which provides equivalent familywise error control. For these reasons, the development of a more powerful pairwise comparison procedure that still controls FWER near α, such as HM or HB, makes it possible and preferable to only conduct all pairwise comparisons. Future research should directly compare the existing one-way *F* tests of equivalence (e.g., Welleck, 2003) with these stepwise procedures to definitively show that conducting controlled pairwise comparisons is sufficiently powerful, while still not allowing differences $\geq \delta$ to be declared equivalent.

While both modified stepwise procedures provide increased power over the CB correction, the HB procedure provides the most power, with rates as high as an uncorrected test in some cases, while still consistently maintaining the Type I error rate below α. However, a Hochberg procedure requires positive dependence of p-values (i.e., when detecting a significant difference (or equivalence) in one pair of means increases the chances of detecting a significant difference (or equivalence) in another; see Lehman, 1966, Benjamini & Yekutieli, 2001). For this reason, Holm-type correction like the HM may be preferable to researchers who cannot guarantee this type of association.

One limitation of the present simulation study is that the conclusions made are based on a finite number of conditions that have been tested. We cannot comment on how these tests will compare under different mean configurations, numbers of groups or sample sizes. However, the conditions for this study were chosen to simulate what is most commonly seen in psychological research. Thus while the results are only applicable to the conditions presented here, we have worked to ensure that the results we collected would reflect what researchers can anticipate to see in their own research as much as possible. Note that to simplify the presentation of the novel methods, we have assumed that all assumptions are satisfied. Since the assumptions of normality and variance homogeneity are regularly violated, we encourage researchers to use robust statistics such as trimmed means with Welch-based test statistics (Cribbie, Fiksenbaum, Wilcox, & Keselman, 2012).

In summary, the present study sought to improve the statistical power of Caffo and colleagues' Bonferroni-type correction of $k^2/4$ when conducting all pairwise comparisons in equivalence testing. By simulating data with a number of different mean configurations, mean sample sizes, sample size configurations and numbers of groups to be compared, we were able to

show that adopting a stepwise procedure (specifically a Holm-type step-down procedure, HM, or a Hochberg-type step-up procedure, HB) provides substantial additional power when the number of pairs of equivalent means is greater than (C - k^2/4), a situation that we believe is common, while still maintaining familywise Type I error rates below α. The results of this study provide psychology researchers with a more powerful tool to assess mean equivalence with three or more groups, and offers an alternative to potentially problematic omnibus tests.

References

Altman, D. G., & Bland, J. M. (1995). Absence of evidence is not evidence of absence. *British Medical Journal, 311*(7003), 485.

Anderson, S.A., & Hauck, W.W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics: Theory and Methods, 12*, 2663–2692.

Beck, A. T., Steer, R. A., Ball, R., & Ranieri, W. F. (1996). Comparison of Beck Depression Inventories-IA and-II in psychiatric outpatients. *Journal of Personality Assessment, 67*(3), 588-597.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics, 29*(4), 1165-1188.

Caffo, B., Lauzon, C., & Röhmel, J. (2013). Correction to "Easy Multiplicity Control in Equivalence Testing Using Two One-Sided Tests". *The American Statistician, 67*(2), 115-116.

Cribbie, R. A. (2016). *equivalencetests*. GitHub repository: https://github.com/cribbie/equivalencetests.

Cribbie, R. A., & Arpin-Cribbie, C. A. (2009). Evaluating clinical significance through equivalence testing: Extending the normative comparisons approach. *Psychotherapy Research, 19*(6), 677-686.

Cribbie, R. A., Arpin-Cribbie, C. A., & Gruman, J. A. (2009). Tests of equivalence for one-way independent groups designs. *The Journal of Experimental Education, 78*(1), 1-13.

Cribbie, R. A., Fiksenbaum, L., Keselman, H. J., & Wilcox, R. R. (2012). Effect of non-normality on test statistics for one-way independent groups designs. *British Journal of Mathematical and Statistical Psychology, 65*(1), 56-73.

Cribbie, R. A., Ragoonanan, C., & Counsell, A. (2016). Testing for negligible interaction: A coherent and robust approach. *British Journal of Mathematical and Statistical Psychology, 69*(2), 159-174.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association. 56* (293): 52–64.

Games, P. A. (1971). Multiple comparisons of means. *American Educational Research Journal, 8*(3), 531-565.

Hancock, G. R., & Klockars, A. J. (1996). The quest for α: Developments in multiple comparison procedures in the quarter century since Games (1971). *Review of Educational Research, 66*(3), 269-306.

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

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics. 6* (2): 65–70.

Kendall, P. C., Marrs-Garcia, A., Nath, S. R., & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology, 67*(3), 285.

Keselman, H. J., Cribbie, R., & Holland, B. (2002). Controlling the rate of Type I error over a large set of statistical tests. *British Journal of Mathematical and Statistical Psychology, 55*(1), 27-39.

Lauzon, C., & Caffo, B. (2009). Easy multiplicity control in equivalence testing using two one-sided tests. *The American Statistician, 63*(2), 147-154.

Lehmann, E. L. (1966). Some concepts of dependence. *The Annals of Mathematical Statistics, 37*(5), 1137-1153.

R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113*, 553–565.

Röhmel, J. (2011). On familywise Type I error control for multiplicity in equivalence trials with three or more treatments. *Biometrical Journal, 53*(6), 914-926.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, 15*, 657–680.

Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods, 3*, 403–411.

Tukey, John (1949). Comparing Individual Means in the Analysis of Variance. *Biometrics. 5* (2): 99–114.

Wellek, S. (2003). *Testing statistical hypotheses of equivalence*. New York: Chapman & Hall/CRC.

Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics*, *32*(4), 741-744.

Appendix A: Tables

Table 1: Adjusted α_{PT} for $\alpha_{FW} = .05$, where $k = 4$, $C = 6$ and $k^2/4 = 4$

Correction Procedure	p-values (smallest – largest)	Correction Factor	α_{PT}
Adjusted Bonferroni (CB; Caffo, Lauzo and Rohmel, 2013)	p_1	4	.0125
	p_2	4	.0125
	p_3	4	.0125
	p_4	4	.0125
	p_5	4	.0125
	p_6	4	.0125
Adjusted Holm (HM)	p_1	4	.0125
	p_2	4	.0125
	p_3	4	.0125
	p_4	3	.0167
	p_5	2	.025
	p_6	1	.05
Adjusted Hochberg (HB)	p_1	4	.0125
	p_2	4	.0125
	p_3	4	.0125
	p_4	3	.0167
	p_5	2	.025
	p_6	1	.05
Traditional Bonferroni (BF; Dunn, 1961)	p_1	6	.0083
	p_2	6	.0083
	p_3	6	.0083
	p_4	6	.0083
	p_5	6	.0083
	p_6	6	.0083
No correction (NC)	p_1	1	.05
	p_2	1	.05
	p_3	1	.05
	p_4	1	.05
	p_5	1	.05
	p_6	1	.05

k = number of groups; n = sample size per cell; α_{PT} = per test α level; α_{FW} = familywise error rate

Table 2: Simulations conditions for $k = 4$, $k = 7$ and $k = 10$

k	Average n	n
4	25	25, 25, 25, 25
		20, 25, 25, 30
		30, 25, 25, 20
	50	50, 50, 50, 50
		40, 50, 50, 60
		60, 50, 50, 40
7	25	25, 25, 25, 25, 25, 25, 25
		19, 21, 23, 25, 27, 29, 31
		31, 29, 27, 25, 23, 21, 19
	50	50, 50, 50, 50, 50, 50, 50
		38, 42, 46, 50, 54, 58, 62
		62, 58, 54, 50, 46, 42, 38
10	25	25, 25, 25, 25, 25, 25, 25, 25, 25, 25
		21, 22, 23, 24, 25, 25, 26, 27, 28, 29
		29, 28, 27, 26, 25, 25, 24, 23, 22, 21
	50	50, 50, 50, 50, 50, 50, 50, 50, 50, 50
		42, 44, 46, 48, 50, 50, 52, 54, 56, 58
		58, 56, 54, 52, 50, 50, 48, 46, 44, 42

k = number of groups; n = sample size per cell

Table 3: Mean configurations for $k = 4$, $k = 7$ and $k = 10$

k	C	$k^2/4$	Population mean configuration
4	6	4	0, 3, 6, 9[a]
			0, 4, 8, 12[a]
			0, 5, 10, 15[a]
			0, 0, 20, 20[b]
			0, 0, 0, 20[c]
			0, 10, 20, 30[c]
			0, 20, 40, 60[d]
7	21	12	0, 1.5, 3, 4.5, 6, 7.5 ,9[a]
			0, 2, 4, 6, 8, 10, 12[a]
			0, 2.5, 5, 7.5, 10, 12.5, 15[a]
			0, 0, 0, 20, 20, 20, 20[b]
			0, 0, 0, 0, 0, 20, 20[c]
			0, 5, 10, 15, 20, 25, 30[c]
			0, 20, 40, 60, 80, 100, 120[d]
10	45	25	0, 1, 2, 3, 4, 5, 6, 7, 8, 9[a]
			0, 1.5, 3, 4.5, 6, 7.5, 9, 10.5, 12, 13.5[a]
			0, 2, 4, 6, 8, 10, 12, 14, 16, 18[a]
			0, 0, 0, 0, 0, 20, 20, 20, 20, 20[b]
			0, 0, 0, 0, 0, 0, 0, 0, 0, 20[c]
			0, 3, 6, 9, 12, 15, 18, 21, 24, 27[c]
			0, 20, 40, 60, 80, 100, 120, 140, 160, 180[d]

C = total number of pairwise comparisons; [a] = pure power; [b] = worse case scenario; [c] = partial power; [d] = pure Type I error

Table 4: Type I error rates for pure Type I error, partial power and worst case scenario mean configurations for $k = 4$

Mean Configuration	n	n config.	CB	HM	HB	BF	NC
Pure Type I Error							
0,20,40,60	25	eq	0.039	0.039	0.039	0.027	0.148
0,20,40,60	25	dc	0.041	0.041	0.041	0.029	0.160
0,20,40,60	25	ac	0.037	0.037	0.037	0.025	0.155
0,20,40,60	50	eq	0.039	0.039	0.039	0.027	0.146
0,20,40,60	50	dc	0.037	0.037	0.037	0.024	0.143
0,20,40,60	50	ac	0.039	0.039	0.039	0.027	0.152
Partial Power							
0,0,0,20	25	ac	0.029	0.032	0.033	0.021	0.112
0,0,0,20	50	eq	0.036	0.046	0.048	0.023	0.118
0,0,0,20	25	eq	0.032	0.035	0.037	0.022	0.110
0,0,0,20	25	dc	0.033	0.037	0.040	0.021	0.113
0,0,0,20	50	dc	0.031	0.039	0.041	0.021	0.108
0,0,0,20	50	ac	0.034	0.043	0.045	0.024	0.117
0,10,20,30	25	eq	0.026	0.026	0.026	0.019	0.101
0,10,20,30	25	dc	0.022	0.022	0.022	0.016	0.098
0,10,20,30	25	ac	0.028	0.028	0.028	0.019	0.100
0,10,20,30	50	eq	0.021	0.021	0.022	0.014	0.096
0,10,20,30	50	dc	0.025	0.026	0.026	0.016	0.098
0,10,20,30	50	ac	0.023	0.023	0.024	0.017	0.096
Worst Case Scenario							
0,0,20,20	25	eq	0.041	0.041	0.041	0.028	0.148
0,0,20,20	25	dc	0.044	0.044	0.044	0.030	0.158
0,0,20,20	25	ac	0.043	0.043	0.043	0.028	0.153
0,0,20,20	50	eq	0.041	0.041	0.041	0.027	0.159
0,0,20,20	50	dc	0.042	0.042	0.043	0.030	0.146
0,0,20,20	50	ac	0.041	0.041	0.042	0.027	0.153

eq = equal sample sizes, ac = ascending sample sizes, dc = descending sample sizes

Table 5: Type I error rates for pure Type I error, partial power and worst case scenario mean configurations for $k = 7$

Mean Configuration	n	n config.	CB	HM	HB	BF	NC
Pure Type I Error							
0,20,…,120	25	eq	0.021	0.021	0.021	0.012	0.278
0,20,…,120	25	dc	0.026	0.026	0.026	0.016	0.274
0,20,…,120	25	ac	0.021	0.021	0.021	0.012	0.275
0,20,…,120	50	eq	0.019	0.019	0.019	0.013	0.269
0,20,…,120	50	dc	0.025	0.025	0.025	0.017	0.275
0,20,…,120	50	ac	0.029	0.029	0.029	0.016	0.274
Partial Power							
0,0,0,0,0,20,20	25	eq	0.034	0.034	0.034	0.020	0.279
0,0,0,0,0,20,20	25	dc	0.030	0.030	0.030	0.019	0.259
0,0,0,0,0,20,20	25	ac	0.033	0.033	0.033	0.023	0.286
0,0,0,0,0,20,20	50	eq	0.030	0.033	0.034	0.018	0.284
0,0,0,0,0,20,20	50	dc	0.030	0.034	0.034	0.017	0.272
0,0,0,0,0,20,20	50	ac	0.037	0.043	0.043	0.024	0.286
0,5,…,30	25	eq	0.011	0.011	0.011	0.007	0.137
0,5,…,30	25	dc	0.014	0.014	0.014	0.008	0.155
0,5,…,30	25	ac	0.009	0.009	0.009	0.004	0.140
0,5,…,30	50	eq	0.012	0.013	0.013	0.007	0.139
0,5,…,30	50	dc	0.010	0.011	0.011	0.005	0.141
0,5,…,30	50	ac	0.011	0.012	0.012	0.007	0.149
Worst Case Scenario							
0,0,0,20,20,20,20	25	eq	0.041	0.044	0.044	0.021	0.322
0,0,0,20,20,20,20	25	dc	0.035	0.035	0.035	0.021	0.331
0,0,0,20,20,20,20	25	ac	0.037	0.037	0.037	0.022	0.311
0,0,0,20,20,20,20	50	eq	0.041	0.042	0.042	0.026	0.328
0,0,0,20,20,20,20	50	dc	0.037	0.037	0.037	0.022	0.316
0,0,0,20,20,20,20	50	ac	0.033	0.034	0.034	0.023	0.313

eq = equal sample sizes, ac = ascending sample sizes, dc = descending sample sizes

Table 6: Type I error rates for pure Type I error, partial power and worst case scenario mean configurations for $k = 10$

Mean Configuration	n	n config.	CB	HM	HB	BF	NC
Pure Type I Error							
0,20,…,180	25	eq	0.015	0.015	0.015	0.009	0.373
0,20,…,180	25	dc	0.014	0.014	0.014	0.006	0.374
0,20,…,180	25	ac	0.018	0.018	0.018	0.010	0.389
0,20,…,180	50	eq	0.019	0.019	0.019	0.008	0.381
0,20,…,180	50	dc	0.019	0.019	0.019	0.012	0.383
0,20,…,180	50	ac	0.019	0.019	0.019	0.012	0.393
Partial Power							
0,0,0,0,0,0,0,0,0,20	25	eq	0.018	0.018	0.018	0.009	0.231
0,0,0,0,0,0,0,0,0,20	25	dc	0.014	0.014	0.014	0.007	0.226
0,0,0,0,0,0,0,0,0,20	25	ac	0.016	0.016	0.016	0.009	0.253
0,0,0,0,0,0,0,0,0,20	50	eq	0.010	0.028	0.029	0.006	0.217
0,0,0,0,0,0,0,0,0,20	50	dc	0.015	0.030	0.030	0.008	0.215
0,0,0,0,0,0,0,0,0,20	50	ac	0.016	0.032	0.033	0.010	0.258
0,3,…,27	25	eq	0.004	0.004	0.004	0.002	0.109
0,3,…,27	25	dc	0.004	0.004	0.004	0.002	0.104
0,3,…,27	25	ac	0.003	0.003	0.003	0.001	0.106
0,3,…,27	50	eq	0.002	0.002	0.002	0.001	0.090
0,3,…,27	50	dc	0.002	0.003	0.003	0.001	0.089
0,3,…,27	50	ac	0.005	0.005	0.006	0.002	0.094
Worst Case Scenario							
0,0,0,0,0,20,20,20,20,20	25	eq	0.040	0.040	0.040	0.023	0.479
0,0,0,0,0,20,20,20,20,20	25	dc	0.036	0.036	0.036	0.019	0.484
0,0,0,0,0,20,20,20,20,20	25	ac	0.039	0.039	0.039	0.021	0.486
0,0,0,0,0,20,20,20,20,20	50	eq	0.042	0.042	0.042	0.024	0.495
0,0,0,0,0,20,20,20,20,20	50	dc	0.035	0.035	0.035	0.020	0.475
0,0,0,0,0,20,20,20,20,20	50	ac	0.039	0.039	0.039	0.022	0.486

eq = equal sample sizes, ac = ascending sample sizes, dc = descending sample sizes

Table 7: Per-pair power rates for pure power, partial power and worst case scenario mean configurations for $k = 4$

Mean Configuration	*n*	*n* config.	CB	HM	HB	BF	NC
Pure Power							
0,3,6,9	25	eq	0.597	0.672	0.706	0.527	0.812
0,3,6,9	25	dc	0.586	0.586	0.696	0.515	0.804
0,3,6,9	25	ac	0.588	0.658	0.691	0.518	0.801
0,3,6,9	50	eq	0.870	0.947	0.953	0.870	0.963
0,3,6,9	50	dc	0.889	0.943	0.949	0.861	0.960
0,3,6,9	50	ac	0.886	0.941	0.947	0.858	0.960
0,4,8,12	25	eq	0.503	0.557	0.587	0.439	0.724
0,4,8,12	25	dc	0.494	0.543	0.574	0.430	0.716
0,4,8,12	25	ac	0.498	0.549	0.578	0.432	0.718
0,4,8,12	50	eq	0.799	0.868	0.877	0.763	0.905
0,4,8,12	50	dc	0.789	0.856	0.865	0.753	0.896
0,4,8,12	50	ac	0.786	0.854	0.865	0.749	0.896
0,5,10,15	25	eq	0.418	0.448	0.469	0.362	0.631
0,5,10,15	25	dc	0.415	0.446	0.464	0.356	0.628
0,5,10,15	25	ac	0.410	0.441	0.460	0.529	0.625
0,5,10,15	50	eq	0.686	0.746	0.756	0.647	0.813
0,5,10,15	50	dc	0.676	0.735	0.746	0.637	0.805
0,5,10,15	50	ac	0.681	0.740	0.750	0.639	0.809
Partial Power							
0,0,0,20	25	eq	0.768	0.769	0.769	0.694	0.931
0,0,0,20	25	dc	0.807	0.807	0.808	0.743	0.951
0,0,0,20	25	ac	0.707	0.708	0.708	0.625	0.907
0,0,0,20	50	eq	0.994	0.995	0.995	0.899	0.999
0,0,0,20	50	dc	0.996	0.996	0.996	0.993	0.999
0,0,0,20	50	ac	0.986	0.986	0.987	0.980	0.999
0,10,20,30	25	eq	0.304	0.304	0.304	0.250	0.539
0,10,20,30	25	dc	0.299	0.299	0.299	0.245	0.537
0,10,20,30	25	ac	0.294	0.294	0.294	0.240	0.534
0,10,20,30	50	eq	0.591	0.591	0.591	0.529	0.802
0,10,20,30	50	dc	0.583	0.584	0.584	0.521	0.797
0,10,20,30	50	ac	0.584	0.585	0.585	0.522	0.792
Worst Case Scenario							
0,0,20,20	25	eq	0.769	0.769	0.769	0.699	0.934
0,0,20,20	25	dc	0.744	0.744	0.744	0.673	0.919
0,0,20,20	25	ac	0.755	0.755	0.755	0.678	0.926
0,0,20,20	50	eq	0.993	0.993	0.993	0.987	0.999
0,0,20,20	50	dc	0.989	0.989	0.989	0.984	0.998
0,0,20,20	50	ac	0.992	0.992	0.992	0.986	0.999

eq = equal sample sizes, ac = ascending sample sizes, dc = descending sample sizes
Note: Power rates for NC tests cannot be compared to other power rates as this procedure was unable to control FWER within a reasonable margin of error of α.

Table 8: Per-pair power rates for pure power, partial power and worst case scenario mean configurations for $k = 7$

Mean Configuration	n	n config.	CB	HM	HB	BF	NC
Pure Power							
0,1.5,…,9	25	eq	0.442	0.487	0.537	0.339	0.846
0,1.5,…,9	25	dc	0.457	0.504	0.557	0.357	0.851
0,1.5,…,9	25	ac	0.426	0.464	0.519	0.325	0.837
0,1.5,…,9	50	eq	0.864	0.944	0.954	0.826	0.976
0,1.5,…,9	50	dc	0.854	0.935	0.947	0.814	0.973
0,1.5,…,9	50	ac	0.850	0.934	0.946	0.808	0.973
0,2,…,12	25	eq	0.386	0.408	0.438	0.294	0.787
0,2,…,12	25	dc	0.397	0.421	0.452	0.308	0.789
0,2,…,12	25	ac	0.368	0.387	0.417	0.277	0.776
0,2,…,12	50	eq	0.779	0.866	0.883	0.734	0.942
0,2,…,12	50	dc	0.763	0.846	0.864	0.716	0.933
0,2,…,12	50	ac	0.762	0.848	0.865	0.716	0.934
0,2.5,…,15	25	eq	0.330	0.339	0.351	0.250	0.710
0,2.5,…,15	25	dc	0.344	0.353	0.367	0.265	0.722
0,2.5,…,15	25	ac	0.319	0.327	0.339	0.241	0.705
0,2.5,…,15	50	eq	0.684	0.751	0.765	0.637	0.881
0,2.5,…,15	50	dc	0.670	0.733	0.748	0.623	0.873
0,2.5,…,15	50	ac	0.672	0.737	0.752	0.623	0.875
Partial Power							
0,0,0,0,0,20,20	25	eq	0.559	0.560	0.560	0.436	0.938
0,0,0,0,0,20,20	25	dc	0.619	0.620	0.620	0.508	0.949
0,0,0,0,0,20,20	25	ac	0.465	0.465	0.465	0.344	0.908
0,0,0,0,0,20,20	50	eq	0.976	0.977	0.977	0.961	0.999
0,0,0,0,0,20,20	50	dc	0.978	0.979	0.980	0.964	0.999
0,0,0,0,0,20,20	50	ac	0.957	0.959	0.959	0.934	0.998
0,5,…,30	25	eq	0.227	0.227	0.227	0.168	0.569
0,5,…,30	25	dc	0.240	0.240	0.240	0.182	0.578
0,5,…,30	25	ac	0.223	0.223	0.223	0.164	0.565
0,5,…,30	50	eq	0.500	0.501	0.501	0.453	0.749
0,5,…,30	50	dc	0.497	0.498	0.498	0.448	0.747
0,5,…,30	50	ac	0.494	0.495	0.495	0.446	0.745
Worst Case Scenario							
0,0,0,20,20,20,20	25	eq	0.552	0.552	0.552	0.425	0.934
0,0,0,20,20,20,20	25	dc	0.547	0.547	0.547	0.429	0.931
0,0,0,20,20,20,20	25	ac	0.557	0.557	0.557	0.446	0.931
0,0,0,20,20,20,20	50	eq	0.977	0.977	0.977	0.962	0.999
0,0,0,20,20,20,20	50	dc	0.961	0.961	0.961	0.939	0.998
0,0,0,20,20,20,20	50	ac	0.970	0.970	0.970	0.953	0.999

eq = equal sample sizes, ac = ascending sample sizes, dc = descending sample sizes
Note: Power rates for NC tests cannot be compared to other power rates as this procedure was unable to control FWER within a reasonable margin of error of α.

Table 9: Per-pair power rates for pure power, partial power and worst case scenario mean configurations for $k = 10$

Mean Configuration	n	n config.	CB	HM	HB	BF	NC
Pure Power							
0,1,…,9	25	eq	0.316	0.324	0.364	0.210	0.859
0,1,…,9	25	dc	0.306	0.313	0.355	0.205	0.855
0,1,…,9	25	ac	0.306	0.314	0.357	0.205	0.856
0,1,…,9	50	eq	0.831	0.927	0.943	0.784	0.981
0,1,…,9	50	dc	0.821	0.920	0.936	0.772	0.979
0,1,…,9	50	ac	0.824	0.922	0.939	0.776	0.980
0,1.5,…,13.5	25	eq	0.256	0.257	0.266	0.168	0.770
0,1.5,…,13.5	25	dc	0.255	0.256	0.265	0.170	0.769
0,1.5,…,13.5	25	ac	0.251	0.252	0.260	0.167	0.765
0,1.5,…,13.5	50	eq	0.698	0.771	0.788	0.646	0.927
0,1.5,…,13.5	50	dc	0.693	0.766	0.782	0.640	0.926
0,1.5,…,13.5	50	ac	0.691	0.763	0.779	0.639	0.924
0,2,…,18	25	eq	0.208	0.208	0.208	0.136	0.672
0,2,…,18	25	dc	0.205	0.205	0.206	0.136	0.668
0,2,…,18	25	ac	0.205	0.205	0.206	0.136	0.671
0,2,…,18	50	eq	0.576	0.606	0.610	0.527	0.836
0,2,…,18	50	dc	0.570	0.599	0.603	0.520	0.832
0,2,…,18	50	ac	0.571	0.600	0.604	0.522	0.833
Partial Power							
0,0,0,0,0,0,0,0,0,20	25	eq	0.388	0.393	0.395	0.260	0.933
0,0,0,0,0,0,0,0,0,20	25	dc	0.404	0.411	0.413	0.279	0.937
0,0,0,0,0,0,0,0,0,20	25	ac	0.352	0.355	0.356	0.232	0.924
0,0,0,0,0,0,0,0,0,20	50	eq	0.956	0.977	0.977	0.931	0.999
0,0,0,0,0,0,0,0,0,20	50	dc	0.957	0.978	0.978	0.933	0.999
0,0,0,0,0,0,0,0,0,20	50	ac	0.943	0.968	0.968	0.913	0.999
0,3,…,27	25	eq	0.165	0.165	0.165	0.107	0.577
0,3,…,27	25	dc	0.162	0.162	0.162	0.107	0.573
0,3,…,27	25	ac	0.162	0.162	0.162	0.107	0.575
0,3,…,27	50	eq	0.462	0.463	0.463	0.416	0.738
0,3,…,27	50	dc	0.459	0.460	0.460	0.413	0.734
0,3,…,27	50	ac	0.459	0.460	0.460	0.414	0.734
Worst Case Scenario							
0,0,0,0,0,20,20,20,20,20	25	eq	0.389	0.389	0.389	0.260	0.935
0,0,0,0,0,20,20,20,20,20	25	dc	0.377	0.377	0.377	0.260	0.926
0,0,0,0,0,20,20,20,20,20	25	ac	0.381	0.381	0.381	0.261	0.928
0,0,0,0,0,20,20,20,20,20	50	eq	0.955	0.956	0.956	0.931	0.999
0,0,0,0,0,20,20,20,20,20	50	dc	0.947	0.947	0.947	0.918	0.999
0,0,0,0,0,20,20,20,20,20	50	ac	0.946	0.946	0.946	0.918	0.999

eq = equal sample sizes, ac = ascending sample sizes, dc = descending sample sizes
Note: Power rates for NC tests cannot be compared to other power rates as this procedure was unable to control FWER within a reasonable margin of error of α.

Table 10: All-pairs power rates for pure power, partial power and worst case scenario mean configurations for $k = 4$

Mean Configuration	*n*	*n* config.	CB	HM	HB	BF	NC
Pure Power							
0,3,6,9	25	eq	0.144	0.361	0.418	0.097	0.418
0,3,6,9	25	dc	0.136	0.345	0.412	0.085	0.412
0,3,6,9	25	ac	0.140	0.349	0.406	0.093	0.406
0,3,6,9	50	eq	0.591	0.807	0.821	0.517	0.821
0,3,6,9	50	dc	0.574	0.796	0.809	0.497	0.809
0,3,6,9	50	ac	0.567	0.794	0.808	0.490	0.808
0,4,8,12	25	eq	0.066	0.207	0.257	0.041	0.257
0,4,8,12	25	dc	0.070	0.193	0.239	0.044	0.239
0,4,8,12	25	ac	0.068	0.199	0.244	0.040	0.244
0,4,8,12	50	eq	0.312	0.563	0.584	0.249	0.584
0,4,8,12	50	dc	0.296	0.537	0.556	0.234	0.556
0,4,8,12	50	ac	0.296	0.535	0.559	0.236	0.559
0,5,10,15	25	eq	0.023	0.095	0.125	0.012	0.125
0,5,10,15	25	dc	0.024	0.095	0.122	0.013	0.122
0,5,10,15	25	ac	0.025	0.093	0.121	0.013	0.121
0,5,10,15	50	eq	0.113	0.281	0.297	0.080	0.297
0,5,10,15	50	dc	0.105	0.260	0.280	0.074	0.280
0,5,10,15	50	ac	0.111	0.275	0.294	0.081	0.294
Partial Power							
0,0,0,20	25	eq	0.542	0.544	0.545	0.435	0.840
0,0,0,20	25	dc	0.602	0.604	0.605	0.499	0.885
0,0,0,20	25	ac	0.456	0.457	0.458	0.352	0.789
0,0,0,20	50	eq	0.984	0.985	0.985	0.972	0.998
0,0,0,20	50	dc	0.989	0.989	0.989	0.981	0.998
0,0,0,20	50	ac	0.962	0.964	0.965	0.948	0.997
0,10,20,30	25	eq	0.003	0.003	0.003	0.001	0.070
0,10,20,30	25	dc	0.002	0.002	0.002	0.000	0.073
0,10,20,30	25	ac	0.001	0.001	0.001	0.002	0.064
0,10,20,30	50	eq	0.116	0.117	0.117	0.065	0.462
0,10,20,30	50	dc	0.104	0.104	0.106	0.057	0.452
0,10,20,30	50	ac	0.109	0.110	0.110	0.062	0.443
Worst Case Scenario							
0,0,20,20	25	eq	0.592	0.592	0.593	0.490	0.873
0,0,20,20	25	dc	0.550	0.550	0.550	0.448	0.843
0,0,20,20	25	ac	0.566	0.566	0.566	0.455	0.857
0,0,20,20	50	eq	0.986	0.987	0.987	0.975	0.998
0,0,20,20	50	dc	0.978	0.979	0.979	0.967	0.996
0,0,20,20	50	ac	0.984	0.984	0.984	0.972	0.997

eq = equal sample sizes, ac = ascending sample sizes, dc = descending sample sizes
Note: Power rates for NC tests cannot be compared to other power rates as this procedure was unable to control FWER within a reasonable margin of error of α.

Table 11: All-pairs power rates for pure power, partial power and worst case scenario mean configurations for $k = 7$

Mean Configuration	n	n config.	CB	HM	HB	BF	NC
Pure Power							
0,1.5,…,9	25	eq	0.002	0.085	0.192	0.000	0.192
0,1.5,…,9	25	dc	0.002	0.091	0.205	0.001	0.205
0,1.5,…,9	25	ac	0.000	0.071	0.179	0.000	0.179
0,1.5,…,9	50	eq	0.199	0.663	0.711	0.135	0.711
0,1.5,…,9	50	dc	0.191	0.640	0.697	0.122	0.697
0,1.5,…,9	50	ac	0.175	0.634	0.692	0.113	0.692
0,2,…,12	25	eq	0.000	0.035	0.095	0.000	0.095
0,2,…,12	25	dc	0.001	0.038	0.098	0.000	0.098
0,2,…,12	25	ac	0.004	0.025	0.087	0.000	0.087
0,2,…,12	50	eq	0.057	0.365	0.433	0.032	0.433
0,2,…,12	50	dc	0.050	0.332	0.399	0.027	0.399
0,2,…,12	50	ac	0.050	0.332	0.399	0.027	0.399
0,2.5,…,15	25	eq	0.000	0.010	0.033	0.000	0.033
0,2.5,…,15	25	dc	0.000	0.010	0.035	0.000	0.035
0,2.5,…,15	25	ac	0.000	0.007	0.029	0.000	0.029
0,2.5,…,15	50	eq	0.009	0.124	0.170	0.005	0.170
0,2.5,…,15	50	dc	0.008	0.107	0.151	0.004	0.151
0,2.5,…,15	50	ac	0.009	0.110	0.154	0.003	0.154
Partial Power							
0,0,0,0,0,20,20	25	eq	0.034	0.042	0.043	0.010	0.627
0,0,0,0,0,20,20	25	dc	0.045	0.053	0.054	0.013	0.656
0,0,0,0,0,20,20	25	ac	0.016	0.020	0.021	0.003	0.516
0,0,0,0,0,20,20	50	eq	0.821	0.835	0.835	0.736	0.991
0,0,0,0,0,20,20	50	dc	0.812	0.828	0.828	0.722	0.986
0,0,0,0,0,20,20	50	ac	0.726	0.745	0.745	0.623	0.979
0,5,…,30	25	eq	0.000	0.000	0.000	0.000	0.000
0,5,…,30	25	dc	0.000	0.000	0.000	0.000	0.001
0,5,…,30	25	ac	0.000	0.000	0.000	0.000	0.000
0,5,…,30	50	eq	0.000	0.000	0.000	0.000	0.006
0,5,…,30	50	dc	0.000	0.000	0.000	0.000	0.006
0,5,…,30	50	ac	0.000	0.000	0.000	0.000	0.005
Worst Case Scenario							
0,0,0,20,20,20,20	25	eq	0.035	0.035	0.035	0.010	0.633
0,0,0,20,20,20,20	25	dc	0.033	0.033	0.033	0.012	0.629
0,0,0,20,20,20,20	25	ac	0.022	0.022	0.022	0.560	0.602
0,0,0,20,20,20,20	50	eq	0.843	0.843	0.843	0.761	0.993
0,0,0,20,20,20,20	50	dc	0.763	0.763	0.763	0.665	0.983
0,0,0,20,20,20,20	50	ac	0.794	0.795	0.795	0.699	0.987

eq = equal sample sizes, ac = ascending sample sizes, dc = descending sample sizes

Note: Power rates for NC tests cannot be compared to other power rates as this procedure was unable to control FWER within a reasonable margin of error of α.

Table 12: All-pairs power rates for pure power, partial power and worst case scenario mean configurations for $k = 10$

Mean Configuration	n	n config.	CB	HM	HB	BF	NC
Pure Power							
0,1,…,9	25	eq	0.000	0.009	0.083	0.000	0.083
0,1,…,9	25	dc	0.000	0.007	0.081	0.000	0.081
0,1,…,9	25	ac	0.000	0.007	0.084	0.000	0.084
0,1,…,9	50	eq	0.045	0.509	0.607	0.022	0.607
0,1,…,9	50	dc	0.040	0.491	0.597	0.017	0.597
0,1,…,9	50	ac	0.043	0.495	0.601	0.020	0.601
0,1.5,…,13.5	25	eq	0.000	0.001	0.018	0.000	0.018
0,1.5,…,13.5	25	dc	0.000	0.001	0.016	0.000	0.016
0,1.5,…,13.5	25	ac	0.000	0.001	0.014	0.000	0.014
0,1.5,…,13.5	50	eq	0.001	0.096	0.174	0.000	0.174
0,1.5,…,13.5	50	dc	0.002	0.095	0.166	0.000	0.166
0,1.5,…,13.5	50	ac	0.001	0.092	0.162	0.001	0.162
0,2,…,18	25	eq	0.000	0.000	0.006	0.000	0.006
0,2,…,18	25	dc	0.000	0.000	0.001	0.000	0.001
0,2,…,18	25	ac	0.000	0.000	0.002	0.000	0.002
0,2,…,18	50	eq	0.000	0.003	0.011	0.000	0.011
0,2,…,18	50	dc	0.000	0.004	0.013	0.000	0.013
0,2,…,18	50	ac	0.000	0.005	0.015	0.000	0.015
Partial Power							
0,0,0,0,0,0,0,0,0,20	25	eq	0.000	0.004	0.005	0.000	0.327
0,0,0,0,0,0,0,0,0,20	25	dc	0.000	0.003	0.005	0.000	0.367
0,0,0,0,0,0,0,0,0,20	25	ac	0.000	0.002	0.003	0.000	0.308
0,0,0,0,0,0,0,0,0,20	50	eq	0.466	0.680	0.682	0.330	0.974
0,0,0,0,0,0,0,0,0,20	50	dc	0.474	0.681	0.683	0.338	0.975
0,0,0,0,0,0,0,0,0,20	50	ac	0.388	0.605	0.607	0.258	0.960
0,3,…,27	25	eq	0.000	0.000	0.000	0.000	0.000
0,3,…,27	25	dc	0.000	0.000	0.000	0.000	0.000
0,3,…,27	25	ac	0.000	0.000	0.000	0.000	0.000
0,3,…,27	50	eq	0.000	0.000	0.000	0.000	0.000
0,3,…,27	50	dc	0.000	0.000	0.000	0.000	0.000
0,3,…,27	50	ac	0.000	0.000	0.000	0.000	0.000
Worst Case Scenario							
0,0,0,0,0,20,20,20,20,20	25	eq	0.000	0.000	0.000	0.000	0.434
0,0,0,0,0,20,20,20,20,20	25	dc	0.000	0.000	0.000	0.000	0.386
0,0,0,0,0,20,20,20,20,20	25	ac	0.000	0.000	0.000	0.000	0.397
0,0,0,0,0,20,20,20,20,20	50	eq	0.546	0.546	0.546	0.404	0.982
0,0,0,0,0,20,20,20,20,20	50	dc	0.490	0.490	0.490	0.346	0.977
0,0,0,0,0,20,20,20,20,20	50	ac	0.480	0.480	0.480	0.345	0.976

eq = equal sample sizes, ac = ascending sample sizes, dc = descending sample sizes

Note: Power rates for NC tests cannot be compared to other power rates as this procedure was unable to control FWER within a reasonable margin of error of α.