A More Powerful Familywise Error Control Procedure for Evaluating Mean Equivalence

Heather Davidson[1] and Robert A. Cribbie[2]

Quantitative Methods Program

Department of Psychology

York University

[1] Department of Psychology, York University, 4700 Keele St., Toronto, Ontario, Canada, M3J 1P3. davi9640@yorku.ca.

[2] Department of Psychology, York University, 4700 Keele St., Toronto, Ontario, Canada, M3J 1P3. cribbie@yorku.ca.

Send correspondence concerning this article to: Robert A. Cribbie, Quantitative Methods Program, Department of Psychology, York University, Toronto, Ontario, Canada, M3J 1P3, cribbie@yorku.ca

Abstract

When one wishes to show no meaningful differences among group means, equivalence tests should be used, as a nonsignificant test of mean difference does not provide evidence supporting equivalence. This research proposes two modified stepwise procedures for controlling the familywise Type I error rate, based on Caffo, Lauzon and Rohmel's (2013) Bonferroni-type correction of $k^2/4$ (where k is the number of groups to be compared). Using a Monte Carlo simulation method, we show that adopting a stepwise procedure increases power, while maintaining the familywise error rate at or below α. Implications for applied research and directions for future study are discussed.

Keywords: Equivalence testing, familywise error, stepwise multiple comparison procedures.

A More Powerful Familywise Error Controlling Procedure for Evaluating Mean Equivalence

Most null hypothesis significance tests are difference-based, meaning that they involve assessing a research hypothesis of a significant relationship, such as a difference between two means, using a null hypothesis of no relationship. However, the results of these tests do not always provide the proper evidence to support researchers' claims. 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 and 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 the 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 a given 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 (Anderson and Hauck 1983; Schuirmann 1987; Westlake 1976). 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, 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.

The two one-sided tests (TOST; Schuirmann 1987) or confidence interval (Westlake, 1976) approach to equivalence testing is the simplest form, 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 for the effect of interest, 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,

Howard and Vessey 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. 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 1-2α 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 applied researchers, equivalence tests have been adapted to fit with different kinds of commonly used statistical tests, including 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. It is important to remember, however, 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 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. 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, and Counsell 2016) as well as difference-based testing (e.g. Games 1971; Hancock and Klockars 1996), and hence we have chosen not to focus on the one-way test of equivalence as a gatekeeper, 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., erroneously rejecting the null hypothesis) 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 be applied by dividing α_{FW} by C such that for each comparison $\alpha_{PT} = \alpha_{FW} / C$, where α_{PT} is the nominal per-test alpha level.

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 and 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 is no difference between two means in the population and from our data we conclude that there is a statistically significant difference, then we make a Type I error – this means that 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. For example, if the true difference in the population has a value of Cohen's *d* equal to three, then it is highly unlikely that we would ever erroneously 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). This means that mean differences that fall close to, but just outside of, the predefined region of equivalence are the most problematic. Researchers (e.g., Rohmel 2011) have defined the area from the border of the equivalence interval up to twice the border of the equivalence interval (i.e., $|\delta| \leq \mu_1 - \mu_1 < |2\delta|$) as a region of problematic Type I errors. As Lauzon and Caffo show, 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 below 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 (see

Figure 1). Only controlling for these so called "problematic" Type I errors provides more power to detect equivalence.

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 $(k - 1)$ provides sufficient Type I error control while resulting in a much less conservative rule than a traditional Bonferroni correction of C. They believed 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 $k \geq 4$. Rohmel showed that a correction of $\alpha_{FW}/ (k - 1)$ (i.e., $\alpha_{FW}/2$) is sufficient to control FWER for $k = 3$, but that a correction of $\alpha_{FW}/4$ is needed to control FWER for $k = 4$, and a correction of $\alpha_{FW}/6$ is needed to control FWER for $k = 5$.

Building on these two studies, Caffo, Lauzon, and Rohmel (2013) proposed a Bonferroni-type correction of $\alpha_{PT} = \alpha_{FW}/(k^2/4)$, where $k^2/4$ 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. For more information on the technical validity of this procedure, see Caffo et al. (2013).

Stepwise Bonferroni Procedures

While Cafflo, Lauzo and Rohmel's procedure has been shown to control FWER at approximately α, research on multiple comparisons in a traditional difference-based framework shows that adopting a stepwise approach makes Bonferroni-type corrections even more powerful while still providing sufficient FWER control (Keselman, Cribbie, and Holland 2002). Holm's step-down Bonferroni procedure (Holm 1979) and Hochberg's step-up Bonferroni procedure (Hochberg 1988) are examples of such procedures. The rationale of these procedures is that they are less conservative by correcting α_{PT} by a smaller factor at each step, while maintaining FWER at or below α.

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, let j be the minimal value where:

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

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.

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:

Commented [RC1]: Please check this change.

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

Researchers reject the null hypotheses $H_j, \ldots, H_C$ and fail to reject the null hypotheses $H_1, \ldots, H_{(j-1)}$. 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.

Both Holm's and Hochberg's sequential Bonferroni procedures will always provide as much or more power than a traditional Bonferroni correction, as *p*-values $P_1, \ldots, P_{C-1}$ are being compared to a less stringent nominal α, and P_C is being compared to the same nominal α. The greatest increases in power can be seen when a number of the null hypotheses are "completely wrong" (Holm 1979); if n of C null hypotheses are completely wrong, these n hypotheses will be easily rejected in the first n steps, leaving the *p*-values that correspond to the remaining $(C - n)$ hypotheses to be evaluated at an $\alpha_{PT} = \alpha_{FW}/(C - n), \ldots, \alpha_{FW}/2, \alpha_{FW}$. Additionally, Hochberg's procedure has the ability to provide more power than Holm's procedure, as its reverse sequential procedure leads researchers to infer significance of *p*-values greater than P_j. Despite their less conservative approaches, both of these procedures have also been shown to maintain FWER at approximately α_{FW}. Taken together, this shows that sequential Bonferroni procedures have the capacity to appropriately control FWER, while not being excessively conservative as are traditional Bonferroni corrections.

Current Study

We propose stepwise multiple comparison procedures for equivalence testing to provide researchers with a more powerful version of Caffo and colleagues' Bonferroni-type correction of

$k^2/4$ (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 will provide more power while still maintaining strong control of α_{FW}.

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 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 only corrects for the potentially problematic Type I error comparisons that fall within the problematic region of $|\delta| \leq \mu_1 - \mu_1 < |2\delta|$. We then test the remaining $k^2/4$ comparisons using Holm's step-down procedure, with $\alpha_{PT} = \alpha_{FW} / (k^2/4 - 1)$, …, $\alpha_{FW} / 2$, α_{FW}. This combination of a stepwise procedure with Caffo and colleagues' maximum correction of $k^2/4$ is expected to provide more power than a simple Bonferroni correction of $k^2/4$, while still providing sufficient Type I error control.

In our adjusted Hochberg procedure (HB), we test the first $C - k^2/4$ comparisons using Hochberg's step-up procedure, with $\alpha_{PT} = \alpha_{FW}, \alpha_{FW} / 2, \ldots, \alpha_{FW} / (k^2/4 - 1)$, 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 is again $k^2/4$, only correcting for comparisons that fall within the problematic region of $|\delta| \leq \mu_1 - \mu_2 < |2\delta|$, 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 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.

Simulation Study

This study used Monte Carlo simulations to evaluate the Type I error rates and power of 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 n = 25 and n = 50, representing typical small and moderate group 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 is 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 $\geq |2\delta|$), 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 (i.e., 0 vs 20, 20 vs 40, etc.). 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

Selected results from the Monte Carlo simulations are presented in Tables 4-7. 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. For full simulation results, please contact the first author.

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 across the conditions studied. 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, as expected, 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 $\alpha = .05$ (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. 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 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 .57 over the CB procedure. The HM procedure produced advantages in all-pairs power of up to .47 over the CB procedure.

Commented [RC2]: See example in next paragraph.

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 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 group 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 one only needs to control for comparisons in which there is 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 unnecessarily controlled for all differences in means, while still maintaining FWER control at or below α.

This study aimed to further increase power, while maintaining FWER closer to α, by utilizing Holm- and Hochberg-type sequential Bonferroni procedures. Our results show 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. Recall that this was shown by Holm (1979), who noted that sequential Bonferroni procedures provide the greatest increase in power when a number of the null hypotheses are "completely wrong". Our modified sequential Bonferroni equivalence testing 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 adjusting α_{PT} *by C, …,* 2, 1 in a Holm-type procedure or by 1, 2, …, C in a Hochberg-type procedure, our procedure involves adjusting α_{PT} by $k^2/4$, …, 2, 1 in the HM procedure and by 1, 2, …, $k^2/4$ 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 (see Tables 4 and 5). 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 (see Tables 6 and 7). 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 follow-up comparisons. That being said, if all pairwise comparisons are significant, the omnibus test is also expected to be significant. 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 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., Wellek 2003) with these stepwise equivalence testing 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 and Yekutieli 2001). For this reason, a Holm-type correction may be preferable for researchers who cannot guarantee this type of association.

Commented [RC3]: Doesn't the direction come into play as well (i.e., positive). E.g., you order the means from smallest to largest. If the smallest mean is really small, then A vs B and A vs C will be POSITIVELY related.

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 applied 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 et al. 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

researchers with a more powerful tool to assess mean equivalence with three or more groups, and offers an alternative to potentially problematic omnibus tests.

Funding details: This work was supported by the Social Sciences and Humanities Council of Canada under Grant 435-2016-1057.

Disclosure statement: The authors declare no conflict of interest.

Bibliography

Altman, D. G., and Bland, J. M. 1995. Absence of evidence is not evidence of absence. *British Medical Journal* 311 (7003), 485. https://doi.org/10.1136/bmj.311.7003.485.

Anderson, S.A., and 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. http://dx.doi.org/10.1080/03610928308828634.

Benjamini, Y., and Yekutieli, D. 2001. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* 29 (4), 1165-1188. http://www.jstor.org/stable/2674075.

Caffo, B., Lauzon, C., and Röhmel, J. 2013. Correction to "Easy Multiplicity Control in Equivalence Testing Using Two One-Sided Tests". *The American Statistician* 67 (2), 115-116. http://dx.doi.org/10.1080/00031305.2012.760487.

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

Cribbie, R. A., Arpin-Cribbie, C. A., and Gruman, J. A. 2009. Tests of equivalence for one-way independent groups designs. *The Journal of Experimental Education* 78 (1), 1-13. http://dx.doi.org/10.1080/00220970903224552.

Cribbie, R. A., Ragoonanan, C., and Counsell, A. 2016. Testing for negligible interaction: A coherent and robust approach. *British Journal of Mathematical and Statistical Psychology* 69 (2), 159-174. http://dx.doi.org/10.1111/bmsp.12066.

Dunn, O. J. 1961. Multiple comparisons among means. *Journal of the American Statistical Association* 56 (293): 52–64. http://dx.doi.org/10.1080/01621459.1961.10482090.

Games, P. A. 1971. Multiple comparisons of means. *American Educational Research Journal* 8 (3), 531-565. https://dx.doi.org/10.3102/00028312008003531.

Hancock, G. R., and 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. https://dx.doi.org/10.3102/00346543066003269.

Hochberg, Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75, 800-802. https://dx.doi.org/10.1093/biomet/75.4.800.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6 (2): 65–70. http://www.jstor.org/stable/4615733.

Kendall, P. C., Marrs-Garcia, A., Nath, S. R., and Sheldrick, R. C. 1999. Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology* 67 (3), 285. http://dx.doi.org/10.1037/0022-006X.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. http://dx.doi.org/10.1348/000711002159680.

Lauzon, C., and Caffo, B. 2009. Easy multiplicity control in equivalence testing using two one-sided tests. *The American Statistician* 63 (2), 147-154. http://dx.doi.org/10.1198/tast.2009.0029.

Lehmann, E. L. 1966. Some concepts of dependence. *The Annals of Mathematical Statistics* 37 (5), 1137-1153. http://dx.doi.org/ 10.1214/aoms/1177699260.

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., and Vessey, J. T. 1993. Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin* 113, 553–565. http://dx.doi.org/10.1037/0033-2909.113.3.553.

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. http://dx.doi.org/10.1002/bimj.201100073.

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. https://dx.doi.org/10.1007/BF01068419.

Seaman, M. A., and Serlin, R. C. 1998. Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods* 3, 403–411. http://dx.doi.org/10.1037/1082-989X.3.4.403.

Tukey, J. 1949. Comparing Individual Means in the Analysis of Variance. *Biometrics* 5 (2): 99–114. http://dx.doi.org/10.2307/3001913.

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. http://dx.doi.org/10.2307/2529259.