Meta-Psychology, 2020, vol 4, MP.2019.1992,
https://doi.org/10.15626/MP.2019.1992
Article type: Original Article

Open data: N/A
Open materials: Yes
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: N/A

Edited by: Erin M. Buchanan
Reviewed by: K.D Valentine, Donald R. Williams
Analysis reproduced by: Erin M. Buchanan
All supplementary files can be accessed at the OSF project page: https://doi.org/10.17605/OSF.IO/9B6Z3

Multiplicity Control vs Replication: Making an Obvious Choice Even More Obvious

Andrew Hunter
York University

Linda Farmus
York University

Nataly Beribisky
York University

Robert Cribbie
York University

This paper presents a side-by-side consideration of multiplicity control procedures and replication as solutions to the problem of multiplicity. Several independent theoretical arguments are presented which demonstrate that replication serves several important functions, and that multiplicity control procedures have a number of serious flaws. Subsequently, the results of a simulation study are provided, showing that under typical conditions, replication provides similar familywise error control and power as multiplicity control procedures. Taken together, these theoretical and statistical arguments lead to the conclusion that researchers who are concerned about the problem of multiplicity should shift their attention away from multiplicity control procedures and towards increased use of replication.

Keywords: multiplicity control, familywise error, power, replication, effect size, meta-analysis

It is easier than ever to collect and analyze vast amounts of data. Plentiful research participants, accessible statistical software, and the popularity of the social sciences have led to a golden age of quantitative research. Much of this research is still being conducted using the lens of Null Hypothesis Significance Testing (NHST). In NHST, tests of "statistical significance" compare the probability of obtaining a test statistic as extreme (or more extreme) than that found under the null hypothesis to a pre-selected nominal Type I error rate. Situations in which findings produced by sampling error are erroneously deemed to be "significant" are referred to as "Type I Errors" or "false positives". As the number of statistical tests being conducted has risen, social science stakeholders have become increasingly concerned with Type I errors (false positives), that is, finding a "significant" effect simply as a result of sampling error. This is because as more and more tests are conducted, the probability of a Type I error occurring increases. Understandably, there have been repeated calls for the adoption of methods (termed "multiplicity control") to reduce the number of false positive results in research (Alibrandi, 2017). At the same time, the value of replication is being touted across many disciplines as a way of ensuring that the results of scientific studies are legitimate (Cumming, 2014; Shrout & Rodgers, 2018).

To date, multiplicity control and replication have rarely been discussed within the same context. This is surprising since they both purport to reduce the likelihood of Type I errors in the results of research studies. Specifically, replications provide more insight over time on the existence (and magnitude) of

effects, while multiplicity control procedures control the rate of decision-making about the existence of given effects within a single framework. In this paper, we discuss the tenets and principles of multiplicity control and replication, and then we move into a comparison of the methods both theoretically and methodologically. We show that one of the many advantages of increased replications is the minimized need for Type I error control via multiplicity control procedures.

Multiplicity Control

Multiplicity refers to testing multiple hypotheses with the goal of isolating those that are statistically significant. The problem is that as the number of tests conducted increases, the probability of obtaining a Type I error also increases. To illustrate this principle, let us say we are comparing the speed at which participants walk. Participants are separated into four groups and each group is primed with a different list of words. If we hypothesize that priming will affect subsequent walking speeds, then we may wish to compare each group to every other group individually (i.e., test all six pairwise comparisons). Though each test carries a specific probability of making a Type I error (α), the overall probability of a Type I error (α') across all six tests will be higher than α. In this way we can see how researchers are often put in the agonizing position of having interesting results that likely contain one or more false positives.

Multiplicity Control Procedures

Researchers have traditionally attempted to control for the increased likelihood of a Type I error when multiple tests are conducted by using multiplicity control procedures (MCPs). There are many different MCPs, but all accomplish essentially the same goal—they make the cut-off demarcating statistically significant from statistically non-significant results more conservative as the number of statistical tests conducted increases (Olejnik, Li, Supattathum, & Huberty, 1997; Sakai, 2018). MCPs can be applied to many kinds of tests, such as pairwise mean comparisons, multiple tests of correlation or regression coefficients, multiple parameters in structural equation modeling, tests repeated over multiple outcome variables, multiple voxel-level tests in fMRI, and more.

Some of the most popular MCPs provide familywise error control (α_{FW}), which controls the probability of at least one Type I error at α across all comparisons (i.e., $\alpha' = \alpha$). The most popular approach for α_{FW} control is the Bonferroni method (Dunn, 1961), which controls for multiplicity by dividing the overall probability of a Type I error (α_{FW}) by the number of tests conducted (T). The resulting per-test alpha level is $\alpha_T = \alpha / T$. Numerous alternatives to the Bonferroni procedure for controlling α' at α have been proposed, such as the Holm (1979) procedure; a flexible and popular alternative. The Holm procedure makes inferences regarding statistical significance in a stepwise manner. The term stepwise implies that the significance tests take place in a prespecified order and α_T can depend on the specific stage of testing.

Replication

Replication lends validity and generalizability to empirical results, and as such, has been heralded as a cornerstone of the so-called "New Statistics" (Cumming, 2014). It also happens that some forms of replication address the multiplicity problem by leveraging the simple principle that it is highly unlikely that sampling error would yield the same false positive result across several studies. These are some of the reasons why replications are gaining traction. Indeed, many academic journals have stated that they are now open to accepting replication studies (e.g., Lucas & Donnellan, 2013; Vazire, 2016). It is our position that replication, an indispensable tool in its own right, naturally and effectively deals with the multiplicity problem.

It is important to note that there are many forms of replication (some have even suggested as many as 12 different types; Radder, 1992). Scholars generally distinguish between direct replications and conceptual replications. Direct replications involve repeating the precise methodology of a previously conducted study, and conceptual replications involve testing the same hypothesis using different methods (Schmidt, 2009). The purpose of a direct replication is to determine the reliability of an effect, whereas a conceptual replication provides a new test of a theory (Simons, 2014).

In the context of the multiplicity problem, direct replications are most relevant, because we are concerned with unreliable effects arising from sampling

error (i.e. Type I errors). If we were concerned with the validity of a claim based on study results (that is, if we wanted to further test whether a given result actually supports a theoretical claim), conceptual replications would be our focus. Therefore, in this paper, when we use the word "replication" we are referring to direct replications.

Multiplicity Control Procedures vs Replication

Although MCPs and replication are theoretically very different, they share a common goal in reducing the probability of Type I errors; hence we find a comparison of these strategies informative. Below we outline the reasons why we find replication to be a more logical and natural way to control for Type I errors than adopting MCPs.

To begin, there are important theoretical issues with the general practice of multiplicity control, highlighted by the fact that there is no basis for the decision to link the alpha (α) level (or maximum acceptable Type I error rate) used for a particular test to the number of other tests conducted within a study (Carmer & Walker, 1985; Cribbie, 2017; Rothman, 1990; Saville, 1990, 2003, 2015). Although many have highlighted that α' can increase drastically when researchers ignore the effects of multiplicity (e.g., Bland & Altma, 1995; Hancock & Klockars, 1996; Holland, Basu & Sun, 2010; Ryan, 1959; Tyler, Normand, & Horton, 2011), there is still no logical theoretical basis for linking the number of tests conducted to the per-test Type I error rate. For example, conducting all T = 6 pairwise comparisons in one study is strategically no different than conducting six studies each with T = 1 pairwise comparison, so why should there be a penalty for conducting all the tests together? If you believe that these situations are equivalent, and that MCPs should be applied in both of these cases, why not control for all tests conducted by the researcher over their scholarly career? Or even further, all statistical tests ever conducted? The ridiculousness of this suggestion speaks to the way that the logic, or lack thereof, underlying MCPs does not scale well. In addition, linking the number of tests conducted to the per-test Type I error rate leads to strange recommendations such as limiting the number of variables studied in order to reduce the potential for Type I errors (Schochet, 2007).

Second, replication involves the repetition of a methodology under slightly different conditions (e.g., different cities, lab settings, research assistants, samples). MCPs only address the likelihood of erroneous results within the study at hand. In contrast, replication reduces error by increasing the scope of an initial study, which directly contributes to the generalizability of the findings (Fisher, 1935; Lindsay & Ehrenberg, 1993). Repeated findings and generalizability—rather than a low chance of error in a single study—have widely been regarded as the hallmark of legitimate results (Carver, 1993; Fisher, 1935; Lykken, 1968; Nuzzo, 2014; Popper, 1934; Steiger, 1990). Methodologists have long stressed the importance of replication for establishing generalizability; for example, Cohen (1994) writes, "For generalization, psychologists must finally rely, as has been done in all the older sciences, on replication." (p. 997).

Third, it has been noted that most (if not all) variables investigated in meaningful studies are related, although the magnitude of the association might not be large (Cohen, 1990, 1994; Gelman, Hill, & Yajima, 2012; Rothman, 2014; Tukey, 1991). This claim suggests that one of the core assumptions of NHST—that the null hypothesis corresponds to a complete non-effect or lack of association—does not map well onto reality. Thus, statistical procedures aimed at reducing Type I errors (like MCPs), which are grounded in NHST, are at best over-conservative, and at worst, unnecessary and irrelevant since Type I errors of this nature are virtually nonexistent. As Tukey (1991) notes, "[it is] foolish…to ask… 'Are the effects [of A and B] different?'…A and B are always different—in some decimal place." (p. 100). And Cohen (1990) states that the null hypothesis can only be true in the "bowels of a computer processor running a Monte Carlo study" (p. 1308). To word this differently, MCPs have the single goal of reducing the likelihood of Type I errors, while replication has the broader goal of exploring the reliability/generalizability of research findings (including the direction and magnitude of effects). If Type I errors do not exist, then MCPs are unnecessary. In contrast, because replication was not specifically designed to address multiplicity (i.e., it is not rooted in NHST, even though some researchers might define replication in terms of NHST results), it remains a valuable pursuit.

Fourth, since replication, unlike MCPs, is not a procedure founded within NHST, there are many

extensions that are available such as focusing on the magnitude of effect sizes and meta-analyzing the effects across the multiple replication studies. In contrast, MCPs are tools which are directly embedded within the NHST framework. Accordingly, MCPs are also subject to the same dichotomous decision-making as the rest of NHST, which has been strongly criticized (e.g., Gigerenzer, Krauss, & Vitouch, 2004).

Lastly, across popular testing situations, multiplicity control is not superior to replication in terms of reducing the likelihood of Type I errors. This novel finding is the primary focus of this paper. If replication is theoretically superior to multiplicity control while providing the statistical benefits of MCPs then there would appear to be a clear winner.

Methodology

We conducted a Monte Carlo simulation to evaluate whether replication provides a similar level of Type I error control to multiplicity control. We simulated pairwise comparisons within a one-way independent groups framework, comparing familywise error control and statistical power of the Bonferroni and Holm MCPs to that of replication (for a list of terms and definitions used in this section, see Table 1).

Several factors were manipulated in this study, including sample size, number of replications, number of populations, population mean configuration, and the method of error control. Sample sizes per group were set at $n = 25$ and $n = 100$ to reflect common sample sizes in psychological research. The number of groups was set at either J = 4 (six pairwise comparisons) or J = 7 (21 pairwise comparisons).

To test both familywise error control and statistical power, we manipulated population means so that three types of configurations were adopted: all population means equal (complete null), some of the population means equal (partial null), and none of the population means equal (complete non-null). In the complete null case, we investigated familywise error rates. In the partial null case, we evaluated both familywise error rates and power. In the complete non-null case, we investigated power. Power was recorded in terms of all-pairs power (the proportion of simulations in which all null hypotheses associated with non-null pairwise comparisons were correctly rejected) and average per-pair power (the proportion of truly significant differences that were correctly identified as statistically significant, averaged across all simulations). The population mean configurations used in the study can be found in Table 2. The within-group standard deviation was fixed at 20, and thus every one-unit increase in differences between means increased the effect size (Cohen's *d*) by .05. For example, the population value for Cohen's *d* for the $\mu_1 = 0$ and $\mu_2 = 16$ comparison is -.8 [(0-16)/20]. In the non-null condition, population means were equally spaced (e.g., 0, 8, 16, 24).

Familywise error control and power were evaluated under situations in which the study was not replicated, replicated once, or replicated twice. With no replication, the familywise error rate was calculated as the proportion of simulations in which at least one pairwise comparison was falsely declared statistically significant (i.e., there was at least one false positive). With one or two replications, the familywise error rate was calculated as the proportion of simulations in which at least one pairwise comparison was falsely declared statistically significant in the original study and in each replication (i.e., the false positive persisted across replications). Note that the order in which the errors are evaluated (study first, replication second; or replication first, study second) is irrelevant since a false effect would need to be present in both to be counted towards the familywise error rate.

With no replication, the per-pair power rate was calculated as the proportion of non-null pairwise comparisons that were correctly declared statistically significant, averaged across all simulations. The all-pairs power rate was the proportion of simulations in which all non-null pairwise comparisons were correctly declared statistically significant. With replication, the per-pair power rate was calculated as the proportion of non-null pairwise comparisons that were correctly declared statistically significant in the original study and in each replication, averaged across all simulations. With replication, the all-pairs power rate was the proportion of simulations in which all non-null pairwise comparisons were correctly declared statistically significant in the original study and in each replication.

To examine familywise error rates across simulations, we adopted three methods of multiplicity control. The first method evaluated each pairwise comparison at α (i.e., no multiplicity control), the second was the Bonferroni MCP method, and the

third was the stepwise Holm MCP method. In addition to computing the test statistics separately for the original study and each replication, to model the accumulation of research over time, Type I error and power rates were also investigated when the combined (meta-analytic) effect across replications was analyzed. Meta-analysis is a useful tool for combining research that examines the same effect, and here, we use it to model how replication effects may be combined.

Lastly, beyond traditional NHST-based approaches, rates were also calculated for instances in which the effect size (Cohen's *d*) meets the minimum meaningful value (ε) in both non-replicated and replicated situations. Since our simulated populations did not violate the assumptions of ANOVA or Cohen's *d*, we did not need to measure amount of bias or use a robust measure of effect size. The effect size equivalent of a Type I error occurs when the observed *d* value mistakenly exceeds ε (i.e., the population value of $d < \varepsilon$, but the observed value of $d > \varepsilon$), whereas the equivalent of power occurs when the observed *d* value correctly exceeds ε (i.e., the population value of $d > \varepsilon$, and the observed value of $d > \varepsilon$). When the population value of $d > \varepsilon$, we can calculate the average proportion of correct statements regarding *d* or the proportion of all correct statements regarding *d* (i.e., all pairwise d values that are greater than ε when population $d > \varepsilon$).

For this study, the nominal Type I error rate was set at $\alpha = .05$, ε was set at $d = .3$, and 5000 simulations were conducted for each condition. It should be noted though that the choice of an appropriate value for α and ε is affected not only by general recommendations but also by the context of the study. Given the lack of context in this study, our choices can be considered somewhat arbitrary.

Finally, it is worth noting that we are simulating perfect, direct replications. As stated earlier, direct replications are useful for 1) testing the reliability of an effect, and 2) establishing the generalizability of an effect. The imperfect nature of replications (i.e., they are conducted in different laboratories, with different kinds of participants, under slightly different conditions) is what makes them useful for establishing generalizability. However, this is not the focus of the present paper. Because this paper is purely interested in the ability of replications to control for multiplicity, we are solely concerned with the way that direct replications can establish the reliability of an effect. Thus, while our simulated replications are artificial and unrealistic, they are all that is needed to compare replication and MCPs in the control of multiplicity.

Results

Familywise Error Control

Tables 3 and 4 show that when studies are not replicated, the Bonferroni and Holm MCPs are, unsurprisingly, vastly preferable to no control. For example, Table 3 shows that in a non-replicated study with 100 participants and 21 comparisons ($\mu = 0, 0, 0, 0, 0, 0, 8$), the Bonferroni and Holm procedures keep the error rate below the nominal $\alpha = .05$. When no control is applied, the familywise error rate greatly exceeds $\alpha = .05$ (.374). In the more extreme situation where no true differences exist ($\mu = 0,0,0,0,0,0,0$), error rates are even worse when no control is used (.440).

When replications are conducted, error rates shift dramatically. Familywise error rates associated with the Bonferroni and Holm MCPs drop to .000, regardless of sample size and number of comparisons. Those associated with no control also drop noticeably. Tables 3 and 4 show that when a single replication is conducted, familywise error rates without multiplicity control are maintained at or below $\alpha = .05$. When two replications are conducted, empirical familywise error rates are maintained below .01.

Although we included results when both multiplicity control and replication are utilized, this paper specifically contrasts the use of multiplicity control in a single (non-replicated) study against a replicated study with no multiplicity control. Thus, the important contrast is between the Bonferroni and Holm results without replication, and that of the no control condition with replication. Tables 3 and 4 show that in this configuration, MCPs and no control with replication both keep empirical familywise error rates at or below $\alpha = .05$. It is important to remind readers that replication was not designed, like MCPs, to maintain familywise error rates at specific levels. In fact, if the number of tests was very large it might take more than one or two replications to control the familywise error rate at α. The statistical properties of replication are nonetheless attractive—the probability of repeating an error over and

over is very slim when the probability of an error in each instance (e.g., study) is small (i.e., α).

Power

A strategy that controls for familywise error by squandering statistical power has little utility. Therefore, we also compared the power provided by MCPs and replication. As expected, overall power (i.e., the probability of finding a significant effect in the original study and each replication) decreases as the number of replications increases, since the probability of finding a statistically significant effect across multiple replications is $(1 - \beta)^R$, where $1 - \beta$ is the power per replication, β is the Type II error rate, and R is the number of replications. Thus, when a partial mean structure is used (μ = 0, 0, 0, 8 or μ = 0, 0, 0, 0, 0, 0, 8), no control with one replication provided similar or higher per-pair power rates than multiplicity control without replication. Per-pair power rates for no control with two replications, and MCPs with no replications, are highly similar. When a complete non-null mean structure was used (μ = 0, 8, 16, 24 or μ = 0, 8, 16, 24, 32, 40, 48) differences in per-pair power between replication and MCPs become slight (often inconsequential).

When all-pairs power was the outcome of interest, results were highly comparable, except when sample sizes were large (e.g., n = 100) and the mean structure contained several true differences (e.g., μ = 0, 8, 16, 24 or μ = 0, 8, 16, 24, 32, 40, 48). In these cases, the Holm procedure often demonstrated superior all-pairs power (see Tables 5 and 6). For example, with population means of 0, 8, 16,24 and n = 100, the Holm procedure had an all-pairs power rate of .450 whereas the all-pairs power rate with no multiplicity control and one replication was .224 and with two replications was .104. The all-pairs power rate for the Bonferroni method was .089.

Meta-Analysis

As meta-analyses are a natural extension of how replications may give more evidence regarding an effect, we also explored the familywise Type I error and power rates when the results of the original study and the replication studies are combined into a single result. As expected, since no multiplicity control is imposed, familywise Type I error rates mirror what would be found in a single study with no multiplicity control. However, since meta-analysis combines the effects of multiple studies, the sample size—and hence the power—rises dramatically. Thus, both the per-pair and all-pairs power rates, especially with larger sample sizes, were much larger than any of the procedures where statistical significance is required in each of the studies conducted. Given recent support for the contention that Type I errors are theoretically implausible in most behavioral science research (e.g., Cribbie, 2017), focusing on power via meta-analytic solutions is very appealing.

Cohen's *d*

Effect sizes have become increasingly popular and are commonly used alongside traditional NHST. Effect size measures allow us to move from the dichotomous determination of the presence or absence of an effect in NHST to an evaluation of magnitudes of effects observed. Unlike statistical significance, measures of effect size do not have statistical cut-offs. Effect size interpretations vary due to both context and magnitude (Beribisky, Davidson, & Cribbie, 2019), However, when there is little theoretical precedence for what constitutes a "meaningful" effect size, unofficial rules of thumb are often used to determine what constitutes a "minimal meaningful value". Here, we have chosen to use an effect size of Cohen's d = .3, which has been conventionally regarded to be within the "small" range of Cohen's d.

Ideally, the choice of an effect may correspond to the smallest meaningful difference (as opposed to how often null comparisons resulted in Cohen's d = 0) since research in many fields (e.g., clinical, health) is often aimed at determining whether observed effects are meaningful or not. Because MCPs are embedded in NHST, which is sample size dependent, MCPs cannot be discussed in relation to observed effect sizes in studies. However, effect sizes can be evaluated across replication studies, and thus we compare our NHST-based results with the ability of replications to prevent a conclusion that the data provides evidence for a meaningful effect size when in fact the population effect size is null. Table 7 summarizes the ability of replication to prevent the effect size equivalent of a Type I error. That is, we present rates at which replication results in a conclusion of a "meaningful" difference between samples

when the population effect size is null. Table 8 summarizes the ability of replication to prevent the effect size equivalent of a Type II error (obtaining a sample effect size that is not meaningful when the population effect size is non-null). Recall that in our simulations we used $d = .3$ as the "minimal meaningful value".

Table 7 shows that when sample sizes are sufficiently large ($n = 100$), replication effectively reduces the frequency of false conclusions about the meaningfulness of Cohen's d. For example, when a partial-null mean structure was used (0, 0, 0, 8 or 0, 0, 0, 0, 0, 0, 0, 8), the probability that a Cohen's d value was erroneously equal to or greater than $d = .3$ across all replications was less than 3% with a single replication, and approximately 0% with two. In other words, the values in Table 7 relate to the proportion of comparisons greater than .3 when the true population difference was 0 (not for conditions when the population Cohen's d is greater in magnitude than zero).

Table 8 reports two different measures. The first is "Average Proportion of Correct" (APC) statements regarding the magnitude of Cohen's d. This is the proportion of Cohen's d values that were accurately at or above $d = .3$ across all simulations. This shows how likely it is that truly meaningful effect sizes will persist across replications. The second is "Proportion of All Correct" (PAC) statements regarding the magnitude of Cohen's d. This is the proportion of simulations in which all truly meaningful effect sizes were sufficiently large ($d \geq .3$) to be labelled as such. This shows how likely it is that every truly meaningful effect size in a study will persist across replications. These two measures are analogous to per-pair power and all-pairs power. Like Table 7, Table 8 reports the proportion of correct statements related to conditions where the effect size in the population is greater than $d = .3$. Table 8 shows that APC and PAC "power rates" for Cohen's d are highly comparable to our previously obtained NHST per-pair and all-pair power rates. Namely, rates decrease as the number of tests increase and the sample sizes decrease. This again highlights the importance of utilizing large sample sizes when possible.

Conclusion

Our simulation study yielded two important conclusions regarding the comparison of multiplicity control and replication on a statistical level. First, both the MCPs and replication maintained Type I error rates at acceptable levels. Second, replication and MCPs provide roughly equivalent power. We also extended the comparison by demonstrating that obvious extensions of replication, such as focusing on effect sizes and meta-analyzing the results of replications, provide valuable research strategies. For example, meta-analysis, as expected, generally provides an advantage in power, although of course at the cost of higher Type I errors without any MCPs. Further, replication is a valuable strategy for minimizing the possibility that a researcher could incorrectly conclude that a meaningful effect size has been detected.

While there were some situations where replication performed better than multiplicity control and vice versa, the overall pattern suggested that MCPs and replication were very similar. Furthermore, our simulations show that when parameters most closely resembled those found in typical social science research studies (i.e., healthy sample size and a moderate number of comparisons where some but not all are truly different), replication provides satisfactory familywise error control and demonstrates equivalent or superior power. Thus, in most situations replication is either as good, or better, than multiplicity control.

Given these results, how should the everyday scientist address the multiplicity problem? It is our position that replication is the best answer. Some may believe this position fails to appreciate practical constraints. Two prominent constraints are limited time and money, and institutional pressure to produce novel (rather than rigorous) results. We recognize the legitimacy and severity of these concerns. However, because of the problematic assumptions underlying MCPs (e.g., null relationships are common), and the subjective nature of many decisions involved in MCPs (e.g., how to define an appropriate "family"), we recommend that they should not be used.

An analogous situation is the use of advanced data analysis techniques (e.g. structural equation modeling, multi-level modeling) by researchers with

limited sample sizes. There may be very good reasons why they cannot access more participants (e.g. lack of access to participants, low funding, time-limited data collection, etc.), and equally valid reasons why their analysis technique would make sense. However, those two truths do not change the fact that their results will be challenging to obtain and interpret with a low sample size. In the same way, the fact that many researchers face barriers to replication does not mean that MCPs are an acceptable answer.

In sum, the results of the present simulation study make the choice to conduct replications, and abandon the use of MCPs, even more obvious by demonstrating that, in addition to being theoretically superior, replication provides natural multiplicity control. Replications also indirectly enable other beneficial research practices such as a comparison of effect sizes and a combining of effect sizes (i.e., meta-analysis). We hope this will encourage members of the social science community to take Wilkinson et al.'s shrewd advice to heart and "let replications promote reputations" (Wilkinson, 1990, p. 600).

Open Science Practices

This article earned the the Open Materials badge for making materials openly available. It has been verified that the analysis reproduced the results presented in the article. The entire editorial process, including the open reviews, are published in the online supplement.

Author Contact

Andrew Hunter
hunter07@yorku.ca
ORCID ID: 0000-0001-7236-0900

Linda Farmus
lifarm@yorku.ca
ORCID ID: 0000-0002-5303-6408

Nataly Beribisky
natalyb1@yorku.ca
ORCID ID: 0000-0002-1081-0125

Robert Cribbie
cribbie@yorku.ca
ORCID ID: 0000-0002-9247-497X

Conflict of Interest and Funding

None of the authors have any conflicts of interest. This research was not funded by any specific source.

Author Contributions

All authors contributed equally to the final paper, including the development of the ideas, the conducting of the simulation study, and the writing of the paper. Robert Cribbie was the senior author and hence is the last author; remaining authorship is alphabetical based on first name.

Appendix

Table 1

Terminology, definitions and more information regarding concepts used in simulation study.

Term	Definition	Idea
n	Sample size per group (25 or 100).	Used to demonstrate commonly found sample sizes within Psychology.
J	Number of groups.	The number of groups directly corresponded to the number of pairwise comparisons such that: Pairwise comparisons = $\frac{J(J-1)}{2}$
Complete Null Condition	All population means are equal.	One of three configurations used for simulation study. Possible to investigate familywise error rates.
Partial Null Condition	Some population means are equal.	One of three configurations used for simulation study. Possible to investigate both familywise error rates and power.
Complete Non-Null Condition	None of the population means are equal.	One of three configurations used for simulation study. Possible to investigate power.
All Pairs Power (AP)	For the true, non-null differences between the groups, all pairwise comparison null hypotheses are correctly rejected.	For: **No replication condition:** Proportion of simulations in which all the real pairwise differences are statistically significant (the null hypothesis is correctly rejected). **One or two replication conditions:** Proportion of simulations in which all non-null pairwise comparisons are statistically significant in the original study and in each replication (the null hypothesis is correctly rejected).
Average Per Pair Power (PP)	For the true, non-null differences between the groups, the corresponding pairwise comparison null hypotheses are correctly rejected (averaged across all simulations).	For: **No replication condition:** Proportion of real pairwise differences that are statistically significant, averaged across all simulations. **One or two replication conditions:** Proportion of real pairwise differences that are statistically significant in the original study and each replication, averaged across all simulations.
Familywise error control	Controls the likelihood of at least one Type I error at α across all comparisons.	For: **No replication condition:** Proportion of simulations where there is at least one pairwise comparison that is incorrectly deemed significant. **One or two replication conditions:** Proportion of simulations where the false positive exists in the original study and the replicated one(s).
Bonferroni method	Type of multiple comparison procedure.	α' = . α/(number of comparisons). *P*-value is compared to α'.
Holm method	Type of multiple comparison procedure.	A sequential modified-Bonferroni procedure that provides greater power while still maintaining strict familywise error control (see Cribbie, 2017 for more details).

d	Cohen's *d* or standardized mean difference.	Type I error: Population *d* is truly less than ε yet *d* incorrectly exceeds ε.
ε	Minimally meaningful value.	Power: population *d* is truly greater than ε and *d* is greater than ε in situation.

Table 2

Simulation Mean Structure

	Familywise Error Control	Familywise Error Control/ Power	Power
6 comparisons	μ = 0,0,0,0	μ = 0,0,0,8	μ = 0,8,16,24
	$\|d_p\| = 0$	$\|d_p\| = 0$ or .4[a]	$\|d_p\|$ = .4, .8, 1.2
21 comparisons	μ = 0,0,0,0,0,0,0	μ = 0,0,0,0,0,0,8	μ = 0,8,16,24,32,40,48
	$\|d_p\| = 0$	$\|d_p\| = 0$ or .40	$\|d_p\|$ = .4,.8,1.2,1.6,2.0 or 2.4

Note. d_p represents the population value for the pairwise Cohen's *d*; when multiple d_p values are provided, e.g., for μ = 0,0,0,8, $|d_p|$ can be 0 or .40, this implies that for some pairwise comparisons $|d_p| = 0$, e.g., for μ_1 vs μ_2, $d_p = |(\mu_1 - \mu_2)/s_p| = |(0-0)/20| = 0$, and for other pairwise comparisons $|d_p| = .40$, e.g., for μ_1 vs μ_4, $d_p = |(\mu_1 - \mu_4)/s_p| = |(0-8)/20| = |-.40| = .40$

Table 3

Familywise Error Rates for 4 groups (T = 6)

	n = 25		n = 100	
	μ = 0,0,0,0	μ = 0,0,0,8	μ = 0,0,0,0	μ = 0,0,0,8
No Replication				
Bonferroni	.041	.023	.038	.022
Holm	.041	.026	.038	.032
No Control	.202	.120	.205	.125
One Replication				
Bonferroni	.001	.000	.000	.000
Holm	.001	.000	.000	.000
No Control	.013	.007	.013	.007
Meta-analysis	.208	.120	.209	.121
Two Replications				
Bonferroni	.000	.000	.000	.000
Holm	.000	.000	.000	.000
No Control	.001	.000	.001	.000
Meta-analysis	.200	.117	.206	.122

Table 4

Familywise Error Rates for 7 groups (T = 21)

	n = 25		n = 100	
	μ = 0,0,0,0,0,0,0	μ = 0,0,0,0,0,0,8	μ = 0,0,0,0,0,0,0	μ = 0,0,0,0,0,0,8
No Replication				
Bonferroni	.039	.028	.043	.028
Holm	.039	.028	.043	.032
No Control	.442	.380	.440	.374
One Replication				
Bonferroni	.000	.000	.000	.000
Holm	.000	.000	.000	.000
No Control	.046	.036	.050	.031
Meta-analysis	.440	.363	.438	.362
Two Replications				
Bonferroni	.000	.000	.000	.000
Holm	.000	.000	.000	.000
No Control	.002	.002	.004	.002
Meta-analysis	.447	.371	.440	.364

Table 5

Average Per-Pair and All Pairs Power Rates for 4 Groups (6 Comparisons)

	n = 25				n = 100			
	μ = 0,0,0,8		μ = 0,8,16,24		μ = 0,0,0,8		μ = 0,8,16,24	
	PP	AP	PP	AP	PP	AP	PP	AP
No Replication								
Bonferroni	.103	.016	.380	.000	.566	.307	.782	.089
Holm	.110	.025	.419	.000	.596	.370	.885	.450
No Control	.288	.092	.567	.002	.802	.610	.903	.470
One Replication								
Bonferroni	.010	.000	.239	.000	.319	.096	.658	.009
Holm	.011	.001	.266	.000	.354	.139	.795	.203
No Control	.080	.008	.410	.000	.647	.374	.825	.224
Meta-analysis	.491	.241	.737	.042	.979	.948	.990	.941
Two Replications								
Bonferroni	.001	.000	.180	.000	.180	.032	.590	.001
Holm	.001	.000	.201	.000	.211	.053	.727	.090
No Control	.024	.000	.332	.000	.524	.233	.762	.104
Meta-analysis	.671	.433	.838	.215	.998	.995	.999	.996

Note. PP = Per Pair power rates, AP = All Pairs power rates

Table 6

Average Per-Pair and All Pairs Power Rates for 7 Groups (21 Comparisons)

	n = 25				n = 100			
	μ = 0,0,0,0,0,0,8		μ = 0,8,16,24,32,40,48		μ = 0,0,0,0,0,0,8		μ = 0,8,16,24,32,40,48	
	PP	AP	PP	AP	PP	AP	PP	AP
No Replication								
Bonferroni	.048	.001	.545	.000	.405	.083	.829	.000
Holm	.050	.002	.594	.000	.420	.100	.913	.150
No Control	.287	.039	.742	.000	.810	.487	.944	.202
One Replication								
Bonferroni	.002	.000	.450	.000	.160	.005	.758	.000
Holm	.002	.000	.494	.000	.172	.008	.851	.021
No Control	.080	.001	.642	.000	.645	.227	.898	.042
Meta-analysis	.501	.147	.907	.000	.977	.903	.994	.874
Two Replications								
Bonferroni	.000	.000	.407	.000	.064	.000	.729	.000
Holm	.000	.000	.448	.000	.071	.001	.809	.002
No Control	.022	.000	.592	.000	.515	.106	.862	.007
Meta-analysis	.682	.312	.890	.031	.998	.990	.999	.990

Note. PP = per pair power rates, AP = all pairs power rates

Table 7

Proportion of Incorrect Statements Regarding the Magnitude of Cohen's d for 4 and 7 Groups

	n = 25		n = 100	
	μ = 0,0,0,0	μ = 0,0,0,8	μ = 0,0,0,0	μ = 0,0,0,8
No Replication	.722	.543	.156	.090
One Replication	.372	.221	.007	.003
Two Replications	.132	.072	.000	.000
	μ = 0,0,0,0,0,0,0	μ = 0,0,0,0,0,0,8	μ = 0,0,0,0,0,0,0	μ = 0,0,0,0,0,0,8
No Replication	.938	.904	.347	.293
One Replication	.714	.612	.025	.016
Two Replications	.358	.275	.002	.000

Table 8

Average Proportion of Correct (APC) *and Proportion of All Correct* (PAC) *Statements Regarding Magnitude of Cohen's d for 4 and 7 Groups*

	n = 25				n = 100			
	μ = 0,0,0,8		μ = 0,8,16,24		μ = 0,0,0,8		μ = 0,8,16,24	
	APC	PAC	APC	PAC	APC	PAC	APC	PAC
No Replication	.651	.404	.810	.173	.758	.543	.880	.368
One Replication	.413	.155	.681	.032	.581	.301	.788	.137
Two Replications	.268	.063	.595	.005	.443	.168	.720	.051
	μ = 0,0,0,0,0,0,8		μ = 0,8,16,24,32,40,48		μ = 0,0,0,0,0,0,8		μ = 0,8,16,24,32,40,48	
No Replication	.647	.268	.890	.016	.765	.422	.931	.117
One Replication	.422	.070	.816	.000	.576	.170	.878	.016
Two Replications	.270	.015	.766	.000	.431	.065	.839	.002

References

Alibrandi, A. (2017). Closed testing procedure for multiplicity control. An application on oxidative stress parameters in Hashimoto's Thyroiditis. Epidemiology, *Biostatistics and Public Health*, 14(1), 1-6. doi: 10.2427/1205

Beribisky, N., Davidson, H., & Cribbie, R. A. (2019). Exploring perceptions of meaningfulness in visual representations of bivariate relationships. *PeerJ*, 7, *e*6853. doi: 10.7717/peerj.6853

Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: the Bonferroni method. *British Medical Journal*, 310(6973), 170. doi: 10.1136/bmj.310.6973.170

Carmer, S. G., & Walker, W. M. (1985). Pairwise multiple comparisons of treatment means in agronomic research. *Journal of Agronomic Education*, 14, 19-26. https://www.crops.org/files/publications/nse/pdfs/jnr014/014-01-0019.pdf

Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education*, 61, 287–292. doi: 10.1080/00220973.1993.10806591

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.

Cohen, J. (1994). The Earth is round (p < .05). *American Psychologist* 49, 997–1003.

Cribbie, R. A. (2017). Multiplicity control, school uniforms, and other perplexing debates. Canadian *Journal of Behavioural Science*, 49, 159-165. doi: 10.1037/cbs0000075

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25(1), 7-29. doi: 10.1177/0956797613504966

Fisher, R. A. (1935). The design of experiments. Oxford, England: Oliver & Boyd

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5(2), 189-211. doi: 10.1080/19345747.2011.618213

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Ed.), *Sage Handbook of Quantitative Methodology for the Social Sciences* (pp. 391–408). Thousand Oaks, CA: Sage.

Hancock, G. R., & Klockars, A. J. (1996). The quest for α: developments in multiple comparison procedures in the quarter century since. *Review of Educational Research*, 66(3), 269-306. doi: 10.3102/00346543066003269

Holland, B., S. Basu, and F. Sun. 2010. Neglect of multiplicity when testing families of related hypotheses. Working Paper, Temple University. doi: 10.2139/ssrn.1466343

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

Lindsay, R. M., & Ehrenberg, A. S. C. (1993). The design of replicated studies. *The American Statistician*, 47(3), 217-228. doi: 10.1080/00031305.1993.10475983

Lucas, R. E., & Donnellan, M. B. (2013). Improving the replicability and reproducibility of research. *Journal of Research in Personality*, 47, 453-454. doi: 10.1016/j.jrp.2013.05.002

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, 70, 151–159. doi: 10.1037/h0026141

Nuzzo, R. (2014). Scientific method: statistical errors. *Nature*, 506(7487), 150-152. doi: 10.1038/506150a

Olejnik, S., Li, J., Supattathum, S., & Huberty, C. J. (1997). Multiple testing and statistical power with modified Bonferroni procedures. *Journal of Educational and Behavioral Statistics*, 22(4), 389-406. doi: 10.3102/10769986022004389

Popper, K. R. (1968). *The logic of scientific discovery*. New York, NY: Harper & Row.

Radder, H. (1992). Experimental reproducibility and the experimenters' regress. In D. Hull, M. Forbes, & K. Okruhlik (Eds.), *Proceedings of the 1992 biennal meeting of the philosophy of science association* (pp. 63–73). East Lansing, MI: Philosophy of Science Association.

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology*, 1(1), 43-46. doi: 10.1097/00001648-199001000-00010

Rothman, K. J. (2014). Six persistent research misconceptions. *Journal of General Internal*

Medicine, *29*(7), 1060-1064. doi: 10.1007/s11606-013-2755-z

Ryan, T.A. (1959). Multiple comparisons in psychological research. *Psychological Bulletin*, 56, 26-47. doi: 10.1037/h0042478

Sakai, T. (2018). *Multiple comparison procedures. In Laboratory experiments in information retrieval*. Springer: Singapore. doi: 10.1007/978-981-13-1199-4_4

Saville, D. J. (1990). Multiple comparison procedures: the practical solution. *The American Statistician*, *44*(2), 174-180. doi: 10.1080/00031305.1990.10475712

Saville, D. J. (2003). Basic statistics and the inconsistency of multiple comparison procedures. *Canadian Journal of Experimental Psychology*, *57*(3), 167. doi: 10.1037/h0087423

Saville, D. J. 2015. Multiple comparison procedures—Cutting the Gordian knot. *Agronomics Journal* (107), 730-735. doi:10.2134/agronj2012.0394.

Schochet, P. Z. (2007). Guidelines for multiple testing in experimental evaluations of educational interventions. Princeton, NJ: Mathematica Policy Research, Inc.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*(2), 90-100. doi: 10.1037/a0015108

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, 69, 487-510. doi: 10.1146/annurev-psych-122216-011845

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76-80.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*(2), 173-180. doi: 10.1207/s15327906mbr2502_4

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, *6*(1), 100-116.

Tyler, K. M., Normand, S. L. T., & Horton, N. J. (2011). The use and abuse of multiple outcomes in randomized controlled depression trials. *Contemporary Clinical Trials*, *32*(2), 299-304. doi: 10.1016/j.cct.2010.12.007

Vazire, S. (2016). Editorial. *Social Psychological and Personality Science*, *7*(1), 3-7. doi: 10.1177/1948550615603955

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594-604. doi: 10.1037/0003-066X.54.8.594