Letting Go of Multiplicity Control: Don't Worry, You Never Liked it Anyway

Rob Cribbie, Nataly Beribisky,
Linda Farmus, & Andrew Hunter

Organization

* Introduction to Multiplicity Control
* Myths regarding Multiplicity Control
* 7 Reasons Why Multiplicity Control is Not Necessary in Modern Psychology Research
* Counter-arguments to Abandoning Multiplicity Control in Psychology
* Recommendations

Introduction to Multiplicity Control

* Whenever we conduct "multiple" tests of significance, we must consider what effect the multiplicity has on the error rates of the test statistics
  * This applies to any type of multiplicity, such as contrasts in ANOVA, predictors in regression, coefficients in SEM, main effects and interactions in factorial models, multiple outcome variables, etc.
  * For example, if we have 4 levels of an IV and we conduct all C = $\binom{k}{2}$ = $k(k\text{-}1)/2 = 4(4\text{-}1)/2 = 6$ pairwise comparisons, each with a Type I error probability of α, what is our *overall* rate of Type I error for the set of comparisons?

Effect of Multiplicity on Type I Errors

* The overall probability of Type I errors approaches $1-(1-\alpha)^C$ for independent comparisons (or approximately $C\alpha$)
  * For example, when we conduct 6 comparisons each at $\alpha = .05$, the overall Type I error rate approaches $1-(1-.05)^6 = .265$ or approximately $6(.05)=.30$
    * Note that these tests are not independent so the overall rate will actually be less than $1-(1-\alpha)^C$, but still higher much higher than α
* The important question is whether or not this rate of Type I error (26.5%) is acceptable

Error Rate Per Test (α_{PT})

* The probability of making a Type I error for any given test
  * For example, we might set α_{PT} to be equal to the nominal α level (e.g., .05)
* However, it becomes evident that as the number of tests to be conducted increases, so does the overall Type I error rate for the set of comparisons

* Is this acceptable??

Familywise Error Rate (α_{FW})

* The probability of making at least one Type I error across a set of tests
* If we maintain $\alpha_{FW} = \alpha$, then α_{PT} will necessarily be $< \alpha$
  * Thus, increasing the number of tests has no effect on the overall Type I error rate, but α_{PT} will become increasingly smaller as the number of comparisons increases

* Is this acceptable??

Example

* Back to our example where we want to conduct all pairwise comparisons with $k = 4$ ($C = 6$) using $\alpha = .05$
  * If we test each of the $C = 6$ pairwise comparisons at $\alpha_{PT} = .05$ then:
    * $\alpha_{fw} \simeq 1-(1-\alpha_{PT})^{C} = 1-(1-.05)^{6} = .265$ (26.5%)
  * If we test each of the $C = 6$ pairwise comparisons at $\alpha_{FW} = .05$ then:
    * $\alpha_{PT} \simeq 1-(1-\alpha_{FW})^{1/C} = 1 - .95^{1/6} = .00852$ (.85%)

Per-test vs Familywise Type I error Control

* So … if you are conducting 6 pairwise comparisons in a one-way ANOVA, which option would you prefer?
  * Per-Test Error Control
    * $\alpha_{PT} = .05$ and $\alpha_{FW} = .265$
  * Familywise Error Control
    * $\alpha_{PT} = .008$ and $\alpha_{FW} = .05$
* Note: The exact same issues apply in other multiplicity testing situations (e.g., 6 outcome variables, 6 path coefficients)

False Discovery Rate Type I Error Control

* The false discovery rate is the expected proportion of false rejections to the total number of rejections
  * This is in contrast to α_{FW}, which is the proportion of one or more false rejections out of the total number of hypothesis tests
* False discovery rate control represents a compromise between strict familywise and liberal per-test control
* FDR is becoming more popular, especially when the number of tests conducted is very large (e.g., fMRI and DNA microarray research)

FDR vs FWE

* Say we do 10 experiments, each with 10 tests (and somehow know which are true/false rejections)
* # of false rejections: 2,1,3,0,0,2,0,1,4,0
* # of total rejections: 7,8,9,6,7,6,7,8,9,6
  * FWE = # of experiments with at least one false rejection over the number of experiments = 6/10 = .6
  * FDR = average proportion of number of false rejections to the total number of rejections
    * FDR = (2/7 + 1/8 + 3/9 + 0 + 0 + 2/6 + 0 + 1/8 + 4/9 + 0)/10 = .165
* Thus, a more liberal (powerful) test would control the FDR at α

Which Type of Control Should You Choose??

* Per-test, Familywise or False Discovery Rate?

* Familywise/False Discover Rate error control have been routinely adopted, recommended, or required by textbook writers, journal editors, etc., and therefore have become the standard in psychological research
  * Even many modern researchers have no patience for anything other than strict multiplicity control

Procedures for Controlling the Familywise Error Rate

* Multiple Planned Comparisons

* Bonferroni
  * α_{PT} is set at α/T, where T represents the number of tests
  * Therefore, if we were to conduct 3 tests and $\alpha_{FW} = .05$, then:
    * $\alpha_{PT} = .05/3 = .0167$

* Extremely conservative if used with correlated tests (e.g., pairwise comparisons, correlation matrices)

Procedures for Controlling the Familywise Error Rate

* Multiple Planned Comparisons or Pairwise Comparisons

* Holm
  * A stepwise Bonferroni procedure that considers the number of possible null hypotheses remaining, given previously rejected null hypotheses
    * The *p*-values are ordered from smallest to largest
      * $p_1, \ldots, p_T$; t = 1, ..., T
    * α_t, starting at t = 1, is set at $\boldsymbol{\alpha}$ / (T-t+1) and if any $p_t > \alpha_t$ testing stops and all remaining p-values (i.e., p_t to p_T) are declared nonsignificant
  * Can be much more powerful than Bonferroni

Procedures for Controlling the False Discovery Rate

* Benjamini-Hochberg Step-Up Procedure
  * Rank the *p*-values (p_t) from smallest to largest ($p_1 \ldots p_T$)
  * α_t, starting at t = T, is set at $\alpha(t/T)$
    * Thus, the largest *p*-value (p_T) is compared against α
      * i.e., (t/T) α = (T/T) α)
  * If any test is significant reject the null for this test and all nulls associated with smaller *p*-values; if any test is not significant go to the next stage of testing

Myths Regarding Multiplicity Control

* Myth 1: Planned Comparisons don't require multiplicity control
  * E.g., Pagano (2013): "With planned comparisons, we do not correct for the higher probability of Type I error that arises due to multiple comparisons, as is done with the post hoc methods … Because planned comparisons do not involve correcting for the higher probability of Type I error, planned comparisons have higher power than post hoc comparisons."
* This argument has no theoretical or mathematical argument
  * Does this mean we can just "plan" to do ALL potential tests?
* Whether you plan or don't plan your tests has no effect on Type I error inflation or the need for multiplicity control

Myths Regarding Multiplicity Control

* Myth 2: Orthogonal Contrasts Don't Require Multiplicity Control
  * Orthogonal contrasts are a set of contrasts that are linearly independent
  * The fact that the SS for the contrasts sums to the SS for the treatment (since the contrasts are independent) has no implications regarding whether or not multiplicity control is necessary

Myths Regarding Multiplicity Control

* Myth 3: Multiplicity control is only required in ANOVA (e.g., pairwise comparisons) but does not apply to multiple correlations, multiple predictors in regression, multiple outcome variables, etc.
  * As already mentioned, this is a ridiculous argument
  * Larzelere & Mulaik (1977) made this argument more than 40 years ago!

Myths Regarding Multiplicity Control

* Myth 4: Multiplicity control is only required if you do lots and lots of tests
  * There is no relationship between the NEED for multiplicity control and the number of tests conducted
    * There is a relationship between the number of tests conducted and the overall probability of a Type I error, but that is a different issue

Myths Regarding Multiplicity Control

* Myth 5: If my tests are correlated, then I don't need multiplicity control

* It is true that the more correlated the test statistics the lower the overall Type I error rate
  * E.g., Imagine two perfectly correlated variables, $\alpha_{FW} = \alpha$
* However, the overall rate of Type I error will exceed α even if a couple moderately correlated tests are conducted
  * And this rate will increase with the number of tests

Myths Regarding Multiplicity Control

* Myth 6: Multiplicity control is not needed if hypotheses are pre-registered
  * This is actually the same argument as planned vs unplanned tests, although since pre-registration is popular right now I decided to make it a separate myth

* Cramer (2018) asserts that if hypotheses are preregistered then the study is confirmatory and the multiplicity issues that affect exploratory analyses do not apply
  * Again, it is not whether the tests are planned or not, it is *how many* tests that are conducted that matters

Do We Really Need Multiplicity Control?

* Earlier it was stated that textbooks, journal editors/policies, and most published articles recommend familywise or false discovery rate Type I error control

* However, the case for NEVER imposing multiplicity control is pretty strong and is gaining momentum
* Next I outline the case AGAINST multiplicity control, via 7 reasons why multiplicity control is unnecessary

Reason 1 : More Power

* If we don't adjust for multiplicity, we have more power for testing our hypotheses

* TRUE … But …
  * Although this is one of the most common reasons provided for dumping multiplicity control, it has no theoretical justification and thus should not be used to justify not adopting multiplicity control
  * The very easy counter-argument is simply to power your study taking into account multiplicity control (Tseng & Shao, 2012)

Reason 2: Simplicity

* Hancock & Klockars (1996)
  * "If [multiplicity control was abandoned], virtually all multiple comparisons would be easily conducted with *t*-tests using liberal critical values, and the MCP researcher would be unemployed."
* Great point … researchers despise having to control for multiplicity, and not just because it lowers power
* Knowing how to define a set of tests over which to impose control, which error rate to control, which procedure to use, how to run the procedure in software, etc. add unwanted time and complexity to the analysis of data
* However, like power, simplicity is not a valid reason for letting go of multiplicity control

Reason 3: Subjectivity in Analysis

* Multiplicity control is, at best, sporadically applied
* Reviewers, editors, etc. are befuddled by when and how multiplicity control should be applied so, in most cases, it is left to the researcher(s) to decide how to apply it
  * And they can find references to support any strategy they prefer
* Thus, researchers are given the potential to make decisions that could affect the conclusions of their study
  * E.g., a researcher analyzes 5 outcomes and three are statistically significant at $\alpha = .05$, but none are significant at $\alpha = .05/5 = .01$
* This subjectivity reinforces the need for more clarity regarding when (if ever) multiplicity control should be imposed, but is not a strong justification for not adopting multiplicity control

Reason 4: Consistency

* Researcher A
  * Explores differences between Arts and Science students on Perfectionism ($T = 1$, $p = .03$)
  * $\alpha = \alpha_{FW} = \alpha_{PT} = .10$ (statistically significant, $p < \alpha_{PT}$)
* Researcher B
  * Explores pairwise differences between Arts, Science, Engineering, Nursing, Health and Humanities students on Perfectionism ($T = 15$, $p_{Arts,Science} = .03$)
    * $\alpha = \alpha_{FW} = .10$
    * $\alpha_{PT} = \alpha_{FW}/T = .10 / 15 = .007$ (Bonferroni)
      * Not statistically significant, $p_{Arts,Science} > \alpha_{PT}$

* These researchers have the same *p*-value, but come to different conclusions regarding the difference between Arts and Science students

Reason 5: The Test of Interest is the Natural Unit of Analysis

* When discussing familywise/false discovery rate error control, we never discussed how you decide upon a *family* of tests

* Imagine a researcher who is evaluating the relationship between each of the Big 5 personality factors and blood flow in 7 brain regions
  * The researcher is going to run this study in three different samples
    * 5-10 year old kids, intro psych students, seniors
  * The researcher is also going to collect blood flow under four conditions
    * Morning/Relaxed, Morning/Stress, Evening/Relaxed, Evening/Stress
* So far the researcher is conducting over 400 tests, and this is Study 1!

Reason 5: The Test of Interest is the Natural Unit of Analysis

* How do we break up these tests into families, in order to impose familywise error control
  * Each brain area is a different family?
    * $\alpha_{PT} = \alpha_{FW}/T = .05/60 = .0008$
  * Each sub-group (kids, etc.) is a different family?
    * $\alpha_{PT} = \alpha_{FW}/T = .05/140 = .0003$
  * Each study is a separate family?
    * Study 1 $\alpha_{PT} = \alpha_{FW}/420 = .05/420 = .0001$
  * Number of tests the researcher conducts this year?
    * $\alpha_{PT} = \alpha_{FW}/T = \alpha_{FW}/? =$ really small!

Reason 5: The Test of Interest is the Natural Unit of Analysis

* Careerwise Type I Error Rate Control
  * O'Keefe (2003), and others, have suggested that if multiplicity control is the standard then what is needed is *careerwise* control, in order to ensure that the number of tests a researcher conducts in his/her career is not related to the probability of a Type I error

* Although absurd, it follows logically from the premise of multiplicity control

Reason 6: There is No Such Thing as a Type I Error

* The whole premise of multiplicity control is that we need to control for situations in which we falsely reject a *true null hypothesis*

* Try to imagine a relationship being investigated in Psychology where the true effect is null
  * E.g., $\rho = 0$, $\mu_1 - \mu_2 = 0$
* If you can, try to imagine a family that contains MULTIPLE null effects

Reason 6: There is No Such Thing as a Type I Error

* Cohen (1990):
  * *"The null hypothesis, taken literally (and that's the only way you can take it in formal hypothesis testing), is always false in the real world. It can only be true in the bowels of a computer processor running a Monte Carlo study (and even then a stray electron may make it false). If it is false, even to a tiny degree, it must be the case that a large enough sample will produce a significant result and lead to its rejection."*

Reason 6: There is No Such Thing as a Type I Error

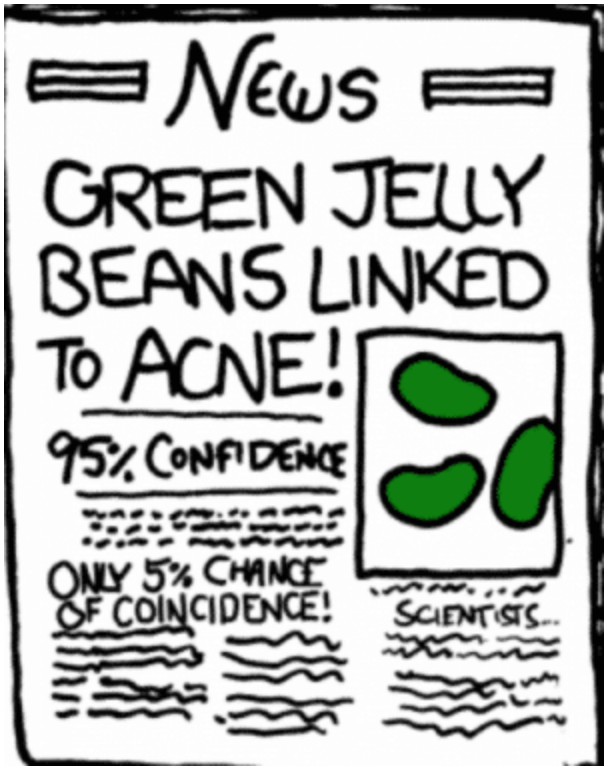
* Meehl (1990)
  * "Everything correlates to some extent with everything else"
  * Meehl also explains that he has found no competent psychologist that disputes this claim

* If there is no such thing as a Type I error, then what on earth are we controlling for?

* Your big worry should be Type II errors … go find more people!

Reason 7: Multiplicity Control is Not a Substitute for Replication

* Multiplicity control was designed to minimize the number of false positives that exist in the research literature
  * However, isn't that the job of replication?
* Replication has solid theoretical support and is one of the key pillars of scientific enquiry
* Would the green jelly beans be linked to acne in a replication?



Reason 7: Multiplicity Control is Not a Substitute for Replication

* Wilkinson and the Task Force on Statistical Inference (1999) had more to say about multiplicity control than any other research design/statistical topic discussed in their now classic paper entitled "Statistical Methods in Psychology Journals: Guidelines and Explanations"

* They conclude the section on multiplicity control by stating:
  * "Let replications promote reputations"

Reason 8: Effect Sizes are the Primary Outcome of Research

* Researchers in the field of Psychology now treat effect sizes (with their accompanying confidence intervals) as the primary outcome of research studies
  * In other words, the focus is on the magnitude of the effects
  * Thus, null hypothesis significance testing now plays a minor role in summarizing effects
* There is no need for multiplicity control in such a framework

Reasons for Not Abandoning Multiplicity Control

* There are two defenses of multiplicity control that are worth discussing
  * Universal null hypothesis
  * Control for multiple tests of single hypotheses, not multiple tests of multiple null hypotheses

Universal Null Hypothesis

* A clinical psychologist is conducting a general mental health check-up on a potential pilot, evaluating their status on depression, anxiety, bipolar disorder, personality disorders, etc.
  * Thus, individual null hypotheses are tested for depression, anxiety, etc., but there is also a *universal* null hypothesis that relates to general mental health
  * Imagine a completely healthy pilot with no mental health issues
    * Any Type I error for an individual hypothesis means a Type I error for the universal null hypothesis

Universal Null Hypothesis

* This sounds like a situation in which it is absolutely necessary to impose multiplicity control
* However, consider the following:
  * Imposing multiplicity control would violate the principle of consistency
    * Rejecting or not rejecting the universal null will depend on how many tests are being conducted on the pilot
  * There is no such thing as a Type I error
    * Find me anyone, not just a pilot, who is perfectly *normal*
      * If you can even define *normal*
  * Replication … test the pilot regularly and look at the 'meta-diagnosis'

Multiple Tests of Multiple Hypotheses vs Multiple Tests of a Single Hypothesis

* Multiple Tests of Multiple Hypotheses (MTMH)
  * Conducting several tests for unrelated (or weakly related) hypotheses
    * E.g., comparing males and females on five different personality constructs (extraversion, agreeableness, etc.)

* Multiple Tests of Single Hypotheses (MTSH)
  * Conducting several tests for the same hypothesis
    * E.g., comparing males and females on extraversion using five different samples

Multiple Tests of Multiple Hypotheses vs Multiple Tests of a Single Hypothesis

* It has been argued that we should impose multiplicity control for MTSH, but not for MTMH (e.g., Matsunga, 2007; Rubin, 2017)

* However, recall:
  * Consistency
    * The authors argue that this imposes consistency, but not in terms of the relationship between α_{PT} and the number of tests
  * The test is the natural unit of analysis
  * There is no such thing as a Type I error
  * Multiplicity control is not a substitute for replication

Going Forward … Some (Expected) Recommendations

* Let go of multiplicity control
  * I promise you won't miss it!
* Base individual effect evidence on effect sizes and confidence intervals for effect sizes
  * *p*-values can be included as supplementary information
* Use meta-analysis to better understand cumulative evidence regarding an effect of interest