Multiplicity Control, School Uniforms, and other Perplexing Debates

Abstract

Researchers in psychology are frequently confronted with the issue of analyzing multiple relationships simultaneously. For example, this could involve multiple outcome variables or multiple predictors in a regression framework. Current recommendations typically steer researchers toward familywise or false discovery rate Type I error control in order to limit the probability of incorrectly rejecting the null hypothesis. Stepwise modified-Bonferroni procedures are suggested for following this recommendation. However, longstanding arguments against multiplicity control, combined with a modern distaste for null hypothesis significance testing, have warranted revisiting this debate. This paper explores both sides of the multiplicity control debate with the goal of educating concerned parties regarding best practices for conducting multiple related tests.

Multiplicity Control, School Uniforms, and other Perplexing Debates

While I was in graduate school, I remember attending a talk where the speaker noted in the introduction that questions surrounding 'Bonferroni or the like' were off limits. As with whether or not a school should adopt uniforms, multiplicity issues are especially perplexing and often create much confusion for researchers. Take, for example, this recent question on ResearchGate: "It seems journals are considering Bonferroni adjustment for p-values of terms within a multiple regression model. Has anyone else noticed this? What do you think of the trend?" (Maltenfort, 2013). This confusion is especially problematic since using multiple outcomes, multiple predictors or multiple treatments in a study both increases the scope of the research and maximizes efficiency (Blakesley et al., 2009). As Wilkinson and the Task Force on Statistical Inference note, "in many areas of psychology, we cannot do research on important problems without encountering multiplicity. We often encounter many variables and many relationships" (1999, p. 599). As a statistical consultant, I have dealt with numerous cases involving multiplicity issues and the one thing that almost all have in common is confusion over which form (if any) of multiplicity control is necessary.

Although these difficult decisions regarding multiple testing fall most directly on the shoulders of the researchers, they are also experienced by journal editors, reviewers, and manuscript readers as they try to interpret the magnitude of the findings of single relationships within a study containing evaluations of multiple relationships. This paper explores the multiplicity issue in psychological research with the goal of making concerned parties more cognizant of the complex issues involved in selecting an appropriate multiple testing strategy.

This paper will highlight two extremely common situations which involve multiple testing. The first is multiple outcomes. It is rare for researchers to conduct a study with only a single primary dependent variable and therefore typically separate tests are conducted on each outcome. For example, a researcher might conduct ten one-way ANOVAs separately on ten different outcome variables, resulting in ten significance tests. The second situation is multiple predictors in regression. Again, with multiple predictors, separate tests on each predictor are almost always conducted. For example, one might explore four different predictors of depression, and statistical tests on each of the coefficients associated with each predictor would be of interest.

Although multiple outcomes and multiple predictors were selected as examples for discussing issues surrounding multiplicity, the same ideas also apply to other multiplicity settings (e.g., voxel-wise activation tests in fMRI). In other words, these two situations are obviously not the only scenarios in which important multiplicity control decisions need to be made, but they provide a platform for introducing the issues involved. In general, any time multiple related of tests of significance are conducted, issues surrounding multiplicity control will arise.

First, multiplicity is defined and the primary issues surrounding multiplicity control are outlined. To do this we revisit one of the most common settings for multiplicity, namely comparing multiple groups on an outcome variable. Second, a discussion of the multiplicity issues involved in each situation is provided. These first two sections are meant to provide the reader with a snapshot of current practice and options for controlling for multiplicity in psychological research. The paper ends by tying together current thinking on multiplicity

control with current thinking on null hypothesis significance testing. The paper aims to raise awareness regarding the issues and potential solutions surrounding multiplicity control and thus provide involved parties with more information on which to make important multiplicity decisions. The overarching goal is to make concrete recommendations regarding the adoption of multiplicity control in psychological research.

Multiplicity/Multiplicity Control

Multiplicity (in the statistical sense) refers to testing multiple hypotheses with the goal of isolating those which are statistically significant. Let's say we are comparing the speed at which participants walk, where participants are separated into four groups. Each group is primed with a different list of words; the word lists related to a race (e.g., winner, fast), fast vehicles (e.g., speedboat, motorcycle), seniors (e.g., grey, cane) or control (e.g., book, chair). The research hypothesis is that priming will affect subsequent walking speeds (see Bargh, Chen & Burrows, 1996). Our interest might be in all of the pairwise comparisons (race vs fast vehicles, race vs seniors, etc.) in order to isolate exactly which differences in priming lists result in differences in walking speed. Thus, we are testing six pairwise null hypotheses, each comparing the population means for two different types of word lists (e.g., H_0: $\mu_{race} = \mu_{fast}$; H_0: $\mu_{race} = \mu_{seniors}$). Each test will be conducted with a specific probability of making a Type I error (α), which is often called a false positive and represents the acceptable probability of concluding that a difference in the population occurs when in fact the population means do not differ.

Now the tricky part: if we conduct multiple tests, each with a Type I error probability α_T (α level for each test), then the overall probability of a Type I error (α') across all six tests will be higher than the probability of a Type I error for each test. If all of the tests are independent, α'

approaches $1 – (1 – \alpha_T)^T$, where T represents the number of tests conducted. It is important to consider how strongly related the tests are that you are conducting, because the strength of the relationships among the tests affects the degree of inflation of α'. Pairwise comparisons are related because some of the test share the same variable (e.g., the tests associated with the hypotheses H_0: $\mu_{race} = \mu_{fast}$ and H_0: $\mu_{race} = \mu_{seniors}$ are related because they share the variable 'rate'; they might also be related because they use a common denominator, but this depends on the nature of the test statistic adopted). As the correlation among the tests increases, α' will be reduced, with $\alpha' = \alpha_T = \alpha$ if the tests were (but realistically speaking never will be) all perfectly correlated (i.e., the decisions are always identical for all tests). In other words, as the correlation among the tests increases, the overall inflation in the Type I error rate decreases. Obviously α' also depends on how many of the null hypotheses are true; for example if none are true then $\alpha' = 0$. [We will dwell on this issue later in the paper.]

For our pairwise comparison problem above, one solution is to treat each test as separate and to ignore the fact that multiple tests of significance are being conducted. In this situation, we are controlling the probability of a Type I error separately for each test, so we can label our multiplicity control strategy 'per test' Type I error control (α_{PT}). This is the simplest form of multiplicity control. As Hancock and Klockars (1996) playfully proposed, if α_{PT} control was unilaterally adopted, all multiple testing would easily be conducted and multiple testing researchers would be unemployed.

A second solution is to control the probability of at least one Type I error at α across all six comparisons (i.e., split up α across all of the tests). Since we are controlling the rate of Type I error over the family of six pairwise hypotheses, this method of Type I error control is often

referred to as familywise error control (α_{FW}). The term *family* refers to the set of tests over which we are concerned about Type I error inflation (e.g., pairwise comparisons, separate tests on each outcome variable, all slope coefficients in a multiple regression). The most popular method of α_{FW} control is the Bonferroni method (Dunn, 1961). This procedure controls for multiplicity by dividing the overall probability of a Type I error α_{FW} by the number of tests conducted. The resulting per-test alpha level is $\alpha_T = \alpha / T$. For a summary of the discussed multiplicity control options and notation, see Table 1. The Bonferroni procedure is an example of a single-step/simultaneous multiple testing procedure since all comparisons are done simultaneously using the same α_T.

Numerous alternatives to the Bonferroni procedure for controlling α' at α have been proposed, with the Holm (1979) procedure a flexible and popular alternative. The Holm procedure makes inferences regarding statistical significance in a stepwise manner. The term *stepwise*, as opposed to single-step/simultaneous, implies that the significance tests take place in a prespecified order and α_T can depend on the specific stage of testing. More specifically, for the Holm procedure the p-values are initially ordered from smallest to largest, and in the first step the smallest p-value is compared against α / T. If this test is not significant, then none of the null hypotheses can be rejected. If this test is significant, then the second smallest p-value is compared to $\alpha / (T - 1)$, again retaining all remaining null hypotheses if the test is not significant and moving on to the next stage of testing if the test is significant. Summarizing, at each stage the *p*-values are compared to $\alpha / (T\text{-}i)$, where *i* represents the number of previously rejected hypotheses; if the *p*-values associated with any hypotheses are not rejected then all the remaining null hypotheses (i.e., those associated with larger *p*-values) are also not rejected.

Lastly, a researcher could also control the expected proportion of the number of Type I errors out of all rejected null hypotheses, called the false discovery rate (α_{FDR}; Benjamini & Hochberg, 1995; Keselman, Cribbie & Holland, 1999). The false discovery rate provides more liberal control of the Type I error rate than α_{FW}, but does not completely ignore the multiplicity issue like α_{PT} (Cribbie, 2003). Note that α_{FDR}-based procedures are more powerful than α_{FW} procedures, in general, because α_{FDR} controls the rate of Type I errors over all rejected hypotheses, instead of all tested hypotheses. Although this is not an exhaustive list of error rates that can be controlled (also see Keselman, Cribbie & Holland, 2002 and Keselman, Miller, & Holland, 2011 for examples of alternative error rates), these three methods of controlling for Type I errors (α_{PT}, α_{FW}, α_{FDR}) represent the most commonly adopted approaches.

Some very convincing arguments for α_{PT} control have been provided by Carmer and Walker (1985), Rothman (1990), Saville (1990; 2003), and others. The primary message of these articles is that conducting T tests of significance within one study has the same goals as conducting T studies each with a single test of significance. In other words, in the pairwise multiple comparison problem above, conducting all T = 6 pairwise comparisons in one study is strategically no different than conducting six studies each with T = 1 pairwise comparisons so why should there be a penalty for conducting all the tests together? Saville takes this argument further and asserts that consistency across studies is important; whether a researcher conducts one test in a study (and therefore does not have to control for multiplicity) or several tests in a study, the number of tests conducted should not affect the conclusion for each individual test. Essentially, the argument is that we should not penalize a researcher for conducting more extensive research. α_{FW} and α_{FDR} procedures are considered inconsistent tests since the α

probability for each test depends on how many other related tests are being conducted. Another argument that has been put forth in support of α_{PT} is that α_{FW} and α_{FDR} increase the probability of a Type II error (e.g., Nakagawa, 2004; Rothman). However, an increased probability of Type II errors should never be used as an argument for adopting α_{PT} control. Decisions regarding multiplicity control should always be made a priori and therefore the study should be powered for the adjusted α level. This point was made clear by Kruschke (2010), who explains how the intentions of the researcher can impact *p*-values, multiple comparisons, etc. As outlined above, different multiple testing procedures lead to different 'levels' of control; if decisions are made a priori, any decisions made after the data is available would need to be treated with skepticism.

However, the arguments in support of α_{PT} have had trouble standing up to the position that conducting multiple tests in a study raises the overall probability of a Type I error, α' (Bland & Altman, 1995; Hancock & Klockars, 1996; Ryan, 1959, 1962; Tyler, Normand, & Horton, 2011). For example, Holland, Basu and Sun (2010) highlighted that false alarms related to the health concerns of Bisphenol A (BPA, in baby bottles), polyvinylchloride (PVC, in toys), and silicone (in breast implants) were caused by researchers ignoring multiplicity.

According to Westfall and Young (1993), if any of the following statements are true then a researcher should be concerned about inflated Type I error rates: 1) It is plausible that some effects might be null; 2) You are prepared to perform many tests of significance in order to find a significant result; 3) Your analysis is exploratory, yet you still want to be confident that a significant result is 'real'; 4) Replication of the experiment is unlikely; 5) There is a "cost" associated with a Type I error; or 6) You want to ensure that the overall rate of Type I error is

held at α. These guidelines cover almost any multiple testing situation imaginable and provide a stark contrast to the arguments of supporters of α_{PT}.

Arguments in favor of α_{FDR} center around the need for a compromise between strict α_{FW} and liberal α_{PT} control, and are especially appealing when the number of hypotheses to be tested becomes very large (e.g., fMRI evaluations where thousands of tests are simultaneously conducted on different voxels within the brain). Cribbie (2003) demonstrated that, in many popular multiple testing situations, α_{FDR} provides an excellent balance between Type I error control and power (i.e., minimizing the combined probability of Type I and Type II errors).

Multiplicity Control with Multiple Outcomes

Few studies in psychology have only a single outcome variable (Lix & Sajobi, 2010; Tyler et al., 2011), and of the majority of studies that have multiple outcome variables, most do nothing to address the multiplicity issue (Vickerstaff, Ambler, King, Nazareth, & Omar, 2015). Wason, Stecher and Mander (2014), Vickerstaff et al., and others have posited that α_{FW} control is necessary when multiple outcomes are analyzed, and Vickerstaff et al. found that if α_{FW} control had been adopted in studies that used no control that the conclusions of many of them would be overturned.

Numerous studies have been conducted that discuss solutions for dealing with multiple outcome variables, exploring a wide variety of options. For example, one could select a single primary outcome, create a composite variable (e.g., sum together the scores), explore an overall treatment effect (e.g., combining the treatment effect across multiple standardized variables), or conduct a multivariate test (Bender & Lange, 2001; Logan & Tamhane, 2004; Wilkinson et al., 1999). Each of these solutions has in common the goal of reducing the testing

problem from multiple tests to a single test, with the multivariate solution the most popular in the literature. If researchers are interested in the multivariate effect and are not interested in the univariate effects of each outcome variable, then multivariate tests are recommended and valuable. However, it is rare that a linear combination of outcome variables is of interest (Huberty & Morris, 1989), and therefore multivariate tests are rarely conducted in practice (Counsell & Harlow, 2016; Vickerstaff et al., 2015). Vickerstaff et al. reported that only two out of sixty trials with multiple primary outcomes used a MANOVA. When they do use a multivariate analysis, they almost always follow-up the test with separate univariate analyses of each outcome (Huberty & Morris; Vickerstaff et al.).

Huberty and Morris reviewed current practice and also explained why the use of a preliminary MANOVA before conducting univariate ANOVAs is not only unnecessary from the standpoint of multiplicity control, but also irrelevant for addressing the dominant research question that relates to the specific treatment effect for each outcome. They reviewed several psychology journals and found that, of researchers who used a MANOVA, only 3 out of 81 conducted a MANOVA without follow-up univariate ANOVAs. Thus, multivariate tests rarely match the theoretical goals of researchers and do little to minimize the effects of multiplicity since univariate tests are a regular follow-up procedure. As Huberty & Morris summarized, “in none of the 222 multiple outcome variable studies was there much interest expressed in any structure associated with the MANOVA results” (p. 302). Lastly, Huberty and Morris, Bird & Hadzi-Pavlovic (2014), and others have shown that using a MANOVA as preliminary test provides only illusory protection of the familywise error rate is

Another solution is to use α_{FW} control (e.g., Bonferroni-type procedures). Although Tyler et al. (2011) found that most researchers did not adjust for multiplicity with multiple outcomes, if they did they used Bonferroni-type procedures. This same finding was reported by Baron, Perrodeau, Boutron, and Ravaud (2013), who found that the original Bonferroni procedure was the most popular adjustment for multiplicity in multi-arm clinical trials.

The original Bonferroni procedure becomes extremely conservative as the number of outcome variables increases, however stepwise modified Bonferroni procedures (e.g., Holm), as discussed previously, provide substantially greater power, good α_{FW} control, and are flexible enough to use in almost any situation (Blakesley et al., 2009). Researchers have explored improvements to the Holm and other stepwise procedures, however rarely are the improvements substantial. For example, Lix and Sajobi (2010) and Troendle (1995) explored resampling-based stepwise procedures that control for the correlations among the outcome variables, however the gains in power were minimal over the original stepwise procedures.

Another option is to weight some outcome variables more highly than others. Alosh, Bretz, and Huque (2011) and Wang et al. (2009) discuss methods based on hierarchically organized hypotheses. In the fixed sequence strategy, all hypotheses are tested at level α but lower order hypotheses are only tested if higher order hypotheses are statistically significant. With fallback procedures, the total α is divided among the hypotheses and rejecting the null hypothesis associated with higher order hypotheses is not required for testing lower order hypotheses. However, if a hypothesis is rejected, all of its weight (α) is carried to the next hypothesis in the sequence. See Alosh et al. and Wang et al. for more details on hierarchical options. Although these procedures provide a better way to spend the allotted α probability

with hierarchically ordered hypotheses, in how many instances are researchers able to weight the importance of their outcome variables? Even more importantly, should the α-level for a given hypothesis depend on how important that outcome variable is relative to other outcome variables?

To conclude, although numerous options are available if α_{FW} (or α_{FDR}) control are desired, stepwise Bonferroni-type procedures (e.g., Holm) are recommended for their Type I error control, power, simplicity and availability in popular software packages (Blakesley et al., 2009). As Dmitrienko and D'Agostino (2013) concluded, unless researchers have an interest in specifying quite different weights to hypotheses, traditional stepwise Bonferroni-type procedures generally perform as well or better than more complicated alternative procedures.

Multiple Regression

Multiple regression is one of the most common forms of analysis conducted in psychology. Two or more variables are used to predict scores on a single outcome variable, and the multiplicity issue arises because researchers are typically interested in the unique contribution of each included predictor. In other words, the statistical significance of each predictor is often of interest and therefore multiple null hypothesis tests are conducted. It is important to point out that if the model is simply used for prediction that no multiplicity issues arise. More specifically, if the full model is used to predict the outcome, without null hypothesis tests for determining which predictors are statistically significant or should be retained in the model, then there are no Type I errors.

Although researchers are generally aware of the inflation in the overall α' with stepwise regression, prescreening of predictors, etc. (Freedman, 1983), it is still rare for researchers to

adopt any form of multiplicity control when assessing the statistical significance of multiple predictors in regression. Larzelere and Mulaik (1977) pointed out about 40 years ago that when assessing the statistical significance of multiple correlations that some sort of multiplicity control is necessary to control for an inflated risk of Type I errors. Cribbie (2000) pointed out the same result when multiple hypotheses are evaluated in structural equation models. The same basic issues apply to multiple regression; if you evaluate the statistical significance of multiple predictors without any form of multiplicity control then α' will be greater than α (with smaller correlations among the predictors resulting in a greater difference between α and α').

As expected, the same solutions proposed for multiple outcomes are applicable for multiple predictors in regression. Stepwise Bonferroni procedures (e.g., Holm) are straightforward, effective at controlling $\alpha_{FW} = \alpha$, available in most statistical software packages and little is gained by adopting more complex versions of these tests (e.g., Troendle, 1995). However, if α_{FDR} control is preferred or different weights are applied to each predictor, then the procedures described in the sections above may provide better results. Smith and Cribbie (2013) also proposed a strategy for controlling for the dependencies among the null hypothesis tests. This approach was proposed within the field of structural equation modeling, however the same procedure is applicable in multiple regression. Essentially, since more highly correlated test statistics lead to less of an inflation of α', Smith and Cribbie proposed that the Bonferroni-type adjustments be weighted by the correlations among the parameters. In the case of multiple regression, this would involve controlling for the correlations among the coefficients for the predictors. Highly correlated parameter estimates would not require the same degree of control as weakly correlated parameter estimates due to the larger amount of

information overlap (recall that α_{FW}, or α' in general, are largest with independent test statistics and no Type I error inflation is observed when the tests are perfectly correlated).

Stop the Bus: The Multiplicity Control and School Uniform Debates are Not Over Yet

Until now, the debate seems pretty one sided. Most of the psychology literature and the discussion above have revolved around how to control α_{FW}. In short, recommendations to date center on the premise that if you want to test multiple hypotheses then you better do something to ensure that there is not a high risk that some tests will be significant by chance alone. However, the marriage of some old ideas about multiplicity control (e.g., Rothman, 1990; Saville, 1990), with a new perspective on the role of null hypothesis testing (e.g., Cumming, 2014), has significantly altered the debate regarding multiplicity control in psychology.

Let's start with revisiting the ideas of Saville (1990, 2003) and Rothman (1990). Saville (1990) notes that the number of Type I errors one makes depends on how many of the tested null hypotheses are true. Since very few (if any) effects are ever null (i.e., very few null hypotheses are ever true), there is little concern about the effects of multiplicity (Gelman, Hill, & Yajima, 2012). Rothman (1990) agrees, commenting that the value of adjustments for multiple testing are seen mostly in simulation studies where random numbers are generated and therefore all null hypotheses are true. Put another way, we assume that the universe is governed by natural laws and relationships, and therefore using the null hypothesis as a starting point is "in effect, to suspend belief in the real world and thereby to question the premises of empiricism" (Rothman, p. 44). To summarize, if the null is never true then there is no such thing as a Type I error and hence controlling for an inflation of the Type I error rate is unnecessary.

A second point raised by Saville (1990) is that the relationship being investigated is the natural unit of analysis, not the family of hypotheses, or the experiment, or the set of experiments, or all hypotheses tested under a research grant, or all hypotheses tested in a researcher's career, or all hypotheses tested in a discipline, etc. This is highly related to the idea of consistency discussed earlier; the criteria for evaluating one relationship should not depend on how many other relationships are investigated.

The next point raised by Saville (1990) liaises nicely with modern thinking (e.g., Cumming, 2014; Gelman et al., 2012). In most research situations, the focus should be on estimation (i.e., precise estimates of parameters), not statistical significance. When the focus is on estimation, rather than statistical significance, multiplicity has no effect on the results or conclusions. In short, null hypothesis tests provide only a small piece of information regarding an association; the bulk of the information comes from model parameters, effect sizes, replication, confidence intervals, meta-analyses, and the like (Cumming; Gelman et al.). The transformation in psychology from a reliance on null hypothesis tests to a distaste for null hypothesis tests, at least in the past few years, has been quick and impactful (Cumming & Calin-Jageman, 2017; Trafimow & Marks, 2015). The quick turnaround in thinking even prompted the American Statistical Association to release a statement on null hypothesis significance testing (Wasserstein & Lazar, 2016), where the authors explain what *p*-values can and cannot tell you about your results.

Imagine that we are exploring the difference between males and females on depression, mathematical ability, and perfectionism. In most situations, our primary focus should be on all facets of the difference between males and females on each outcome variable, what Tukey

(1977) called exploratory data analysis, not only on the statistical significance of each of these comparisons or controlling for how many tests we are conducting. This sort of analysis includes mean differences, median differences, differences in variability, distributional shape differences, differences in outliers, posterior distributions, etc. This would also include ensuring precise estimation of the differences via large sample sizes, valid and reliable instruments, proper models, and encouraging replications of the findings. For example, Gelman et al. (2012) make it clear that by adopting more sophisticated models (e.g., multilevel models with Bayesian estimates) we can obtain more robust estimates of the parameters. At the same time, as Cohen (in press) highlights, it is important to acknowledge that different areas of research have different goals, and the focus of research in one area, say effect sizes, might be quite different from the focus in another area, e.g., direction of effects. In general though, it is important to think of each relationship as the natural unit of analysis and to focus all available energy on finding out the most about that relationship as possible.

Can null hypothesis testing be a part of this process? Currently, that is a much debated point, with some completely opposed to *p*-values and null hypothesis testing (e.g., Trafimow & Marks, 2015; Cumming, 2014), while others are more tolerant and recognize the value of the information provided by the *p*-value and its interpretation (Wasserstein & Lazar, 2016). There is little debate that the focus of most current research in psychology is still typically hypothesis testing, however methodological discussions at conferences, in blogs and papers, etc. seem to indicate that more radical changes to practice are near. Obviously this observation is subjective, however in addition to the published articles supporting this contention referred to above, there are also some quantitative indicators that discussions surrounding null hypothesis testing

are spiking. For example, a 'Google Scholar' search for the words "null", "hypothesis" and "testing" in the title of articles returned 16 hits for articles published in 2005/2006, 27 hits for articles published in 2010/2011, and 49 hits for articles published in 2015/2016.

If null hypothesis testing becomes only a minor part of the research process, then should multiplicity control be necessary? There is no debating that α' will rise as the number of tests increases, but is that an issue when our focus is on precisely estimating the nature of the relationship? We also cannot forget the convincing arguments of Saville (1990) and Rothman (1990) for not invoking multiplicity control at all.

Conclusion

Some debates continue for years with no sign of resolve, including school uniforms, free will, nature/nurture, and whether the chicken or the egg came first. The debate over school uniforms has continued for decades. For example, proponents see a valuable way to minimize arguments/bullying over fashion/style, while opponents see the loss of a child's sense of individuality. Interestingly, proponents of both sides argue that costs are lower with their preference. Is the multiplicity debate as unwinnable as the school uniform debate? For many years the prevailing sentiment has been that it is important to control for multiplicity when conducting multiple null hypothesis tests, and this sentiment has not only applied to multiple comparisons in an ANOVA setting but also to regression, multiple outcome variables, structural equation modeling, voxel evaluations in functional magnetic resonance imaging, etc.

However, the recent abrupt shift in focus from null hypothesis testing to precise estimation of relationships (at least, for now, in the methodological literature) has definitely made the debate more perplexing (and interesting). The role of multiplicity control in

psychology will depend to a large extent on the role that null hypothesis testing and dichotomous thinking play in psychology. Currently, null hypothesis testing still dominates published research and instruction in psychology and if that continues researchers need to be aware of and use the techniques described earlier in the paper because they will likely be required by book/journal editors, reviewers, etc. Editors and reviewers list conducting multiple tests without adjustment as one of the most popular statistical errors in submitted manuscripts (Harris, Reeder, & Hyun, 2011). On the other hand, if, as expected, null hypothesis testing eventually takes on a very minor role in the evaluation of relationships, then discussions surrounding multiplicity control may all but disappear. Precisely estimating relationships avoids dichotomous thinking and errors of statistical inference (Type I/II), and thus multiplicity control is unnecessary.

Here is a brief summary regarding recommendations for evaluating multiple relationships, be that multiple comparisons of means, multiple outcomes, multiple predictors in regression, etc.:

1) If the statistical analyses rely heavily on null hypothesis significance testing (and it should be clear from the discussion above that this strategy is generally not recommended), then the nature/goals of the research might drive the decision regarding multiplicity control. However, policies of the publication outlet for the research might also come into play; although researchers might agree with proponents of α_{PT} control, currently most editors and reviewers require α_{FW} (or in some cases α_{FDR}) control in order to minimize the probability of Type I errors. If α_{FW} (or α_{FDR}) multiplicity control is necessary, stepwise Bonferroni-type procedures (e.g.,

Holm) are recommended for their simplicity, balance of Type I error control and power, and availability in popular software packages such as R, SAS and STATA.

2) If the statistical analyses prioritize estimation over null hypothesis significance testing, the recommended approach in most research settings, then multiplicity control is unnecessary since hypothesis testing (and hence Type I errors) are a minor part of the analysis strategy.

Like a parent waiting on a decision regarding school uniforms, the field of psychology is currently waiting to see what position editors, textbook authors, governing bodies, etc. take regarding null hypothesis significance testing and, consequently, multiplicity control. However, Gigerenzer (2004), regarding ritualistic null hypothesis testing, recommends a more active approach that definitely applies here: “We need some pounds of courage to cease playing along in this embarrassing game. This may cause friction with editors and colleagues, but it will in the end help them to enter the dawn of statistical thinking” (p. 604). Current sentiments point towards *p*-values and null hypothesis testing taking on a very minor role in quantitative methods for psychology, but for now the debate rages on and the decisions, approaches and rationales of individual researchers may determine the future path of multiple testing analyses in psychology.

References

Alosh, M., Bretz, F. & Huque, M. (2014). Advanced multiplicity adjustment methods in clinical trials. *Statistics in Medicine, 33*, 693–713. DOI:10.1002/sim.5974.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71*, 230-244. DOI: http://dx.doi.org/10.1037/0022-3514.71.2.230.

Baron, G., Perrodeau, E., Boutron, I., & Ravaud, P. (2013). Reporting of analyses from randomized controlled trials with multiple arms: A systematic review. *BMC medicine*, *11*, 1. DOI: 10.1186/1741-7015-11-84.

Bender, R. & Lange, S. (2001). Adjusting for multiple testing—when and how? *Journal of Clinical Epidemiology, 54*, 343-349. DOI: http://dx.doi.org/10.1016/S0895-4356(00)00314-0.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B, 57*, 289-300. DOI: http://www.jstor.org/stable/2346101.

Blakesley, R. E., Mazumdar, S., Amanda Dew, M., Houck, P. R., Tang, G., Reynolds III, C. F., Butters, M. A. (2009). Comparisons of methods for multiple hypothesis testing in neuropsychological research. *Neuropsychology, 23*, 255-264. DOI: 10.1037/a0012850.

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

Cohen, B. (in press). Why the resistance to statistical innovations? A Comment on Sharpe (2013). *Psychological Methods*. DOI: 10.1037/met0000058.

Counsell, A. & Harlow, L. L. (in press). Reporting practices and use of quantitative methods in Canadian journal articles in psychology. *Canadian Psychology*. http://dx.doi.org/10.1037/cap0000074

Cribbie, R. A. (2000). Evaluating the importance of individual parameters in structural equation modeling: The need for Type I error control. *Personality and Individual Differences, 29*, 567-577. DOI: http://dx.doi.org/10.1016/S0191-8869(99)00219-6.

Cribbie, R. A. (2003). Pairwise multiple comparisons: New yardstick, new results. *Journal of Experimental Education, 71,* 251-265. DOI: http://www.jstor.org/stable/20152711.

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

Cumming, G., & Calin-Jageman, R. (2017). *Introduction to the new statistics: Estimation, open science, and beyond.* New York: Routledge.

Dmitrienko, A. & D'Agostino, R. (2013). Traditional multiplicity adjustment methods in clinical trials. *Statistics in Medicine, 32*, 5172-5218. DOI: 10.1002/sim.5990.

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

Freedman, D. A. (1983). A note on screening regression equations. *The American Statistician, 37*, 152-155. DOI:10.1080/00031305.1983.10482729.

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*, 189–211. DOI: 10.1080/19345747.2011.618213.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*, 587-606.

Hancock, G. R., & Klockars, A. J. (1996). The quest for α: Developments in multiple comparison procedures in the quarter century since Games (1971). *Review of Educational Research, 66*, 269-306. DOI: 10.3102/00346543066003269.

Harris, A., Reeder, R., & Hyun, J. (2011). Survey of editors and reviewers of high-impact psychology journals: Statistical and research design problems in submitted manuscripts. *The Journal of Psychology: Interdisciplinary and Applied, 145*, 195-209. DOI:10.1080/00223980.2011.555431.

Holland, B., Basu, S. & Sun, F. (2010). *Neglect of multiplicity when testing families of related hypotheses*. Working paper, Temple University.

Holland, B., & Cheung, S. H. (2002). Familywise robustness criteria for multiple-comparison procedures. *Journal of Royal Statistical Society B, 64*, 63-77.

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

Huberty, C. J., & Morris, J. D. (1989). Multivariate analysis versus multiple univariate analyses. *Psychological Bulletin, 105*, 302-308. DOI: http://dx.doi.org/10.1037/0033-2909.105.2.302.

Keselman, H. J., Cribbie, R., & Holland, B. (1999). The pairwise multiple comparison multiplicity problem: An alternative approach to familywise/comparisonwise Type I error control. *Psychological Methods, 4*, 58-69. DOI: 10.1037/1082-989X.4.1.58.

Keselman, H. J., Cribbie, R. A. & 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*, 27-39. DOI: 10.1348/000711002159680.

Keselman, H. J., Miller, C. W., & Holland, B. (2011). Many tests of significance: New methods for controlling Type I errors. *Psychological Methods, 16*, 420-31. DOI: 10.1037/a0025810.

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*, 658-676. http://dx.doi.org/10.1002/wcs.72.

Larzelere, R. E., & Mulaik, S. A. (1977). Single sample tests for many correlations. *Psychological Bulletin, 84*, 557-569. DOI: http://dx.doi.org/10.1037/0033-2909.84.3.557.

Lix, L. M. & Sabjobi, T. (2010). Testing multiple outcomes in repeated measures designs. *Psychological Methods, 15*, 268–280. DOI: 10.1037/a0017737.

Logan, B. R., & Tamhane, A. C. (2004). On O'Brien's OLS and GLS tests for multiple endpoints. *Lecture Notes-Monograph Series, Recent Developments in Multiple Comparison Procedures, 47*, 76-88. DOI: 10.1214/lnms/1196285627.

Maltenfort, M. (Jan. 7, 2013). *Bonferroni correction for multiple regression models? - ResearchGate*. Available from: https://www.researchgate.net/post/Bonferroni_correction_for_multiple_regression_models [accessed Jul 13, 2016].

Marcus, R., Peritz, E., & Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika, 63*, 655–660. DOI: 10.1093/biomet/63.3.655.

Nakagawa, S. (2004). A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology, 15*, 1044-1045. DOI: https://doi.org/10.1093/beheco/arh107.

Rothman,

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

Ryan, T. A. (1962). The experiment as the unit for computing rates of error. *Psychological Bulletin, 59*, 301-305. DOI: http://dx.doi.org/10.1037/h0040562.

Saville, D. J. (1990). Multiple comparison procedures: The practical solution. *The American Statistician, 44*, 174- 180. DOI: 10.2307/2684163.

Saville, D. J. (2003). Basic statistics and the inconsistency of multiple comparison procedures. *Canadian Journal of Experimental Psychology, 57*, 167-175. http://dx.doi.org/10.1037/h0087423.

Smith, C., & Cribbie, R. A. (2013). Significance testing in structural equation modeling: Incorporating parameter dependencies into multiplicity controlling procedures. *Structural Equation Modeling: An Interdisciplinary Journal*, *20*, 79-85. DOI: DOI:10.1080/10705511.2013.742385.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37,* 1-2, DOI: 10.1080/01973533.2015.1012991.

Troendle, J. F. (1995). A stepwise resampling method of multiple hypothesis testing. *Journal of the American Statistical Association, 90*, 370-378. DOI: 10.1080/01621459.1995.10476522.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

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

Vickerstaff, V., Ambler, G., King, M., Nazareth, I., & Omar, R. Z. (2015). Are multiple primary outcomes analysed appropriately in randomized controlled trials? A review. *Contemporary Clinical Trials, 45*, 8–12. DOI: 10.1016/j.cct.2015.07.016.

Wang, D., Li, Y., Wang, X., Liu, X, Fu, B., Lin, Y., Larsen, L., & Offen, W. (2009). Overview of multiple testing methodology and recent development in clinical trials. *Contemporary Clinical Trials, 45*, 13-20. DOI: http://dx.doi.org/10.1016/j.cct.2015.07.014.

Wason, J. M. S., Stecher, L., & Mander, A. P. (2014). Correcting for multiple-testing in multi-arm trials: Is it necessary and is it done? *Trials, 15*, 1-7. DOI: 10.1186/1745-6215-15-364.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Statement on p-Values: Context, process, and purpose. *The American Statistician, 70*, 129-133. DOI: 10.1080/00031305.2016.1154108.

Westfall, P.H. & Young, S.S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. New York: Wiley.

Wilkinson, L. & The Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604. DOI: http://dx.doi.org/10.1037/0003-066X.54.8.594.

Table 1.

Summary of the notation regarding per-test and familywise multiplicity control.

Type of Multiplicity Control	**Proposed Overall Type I Error Rate (α')**	**Type I Error Rate for Each Test (α_T)**
Per-Test (α_{PT})	$\alpha' \leq 1 - (1 - \alpha_T)^T$	$\alpha_T = \alpha$
Familywise (α_{FW})	$\alpha' = \alpha$	$\alpha_T \leq \alpha$ (depends on procedure) e.g., Bonferroni $\alpha_T = \alpha / T$

Note: α = maximum permissible Type I error rate; α_{PT} = per-test α level; α_{FW} = familywise Type I error rate; T = number of statistical tests.