Title
Type I Error Control in Psychology Research: Improving Understanding in General and Addressing Multiplicity in Some Specific Contexts

Author
Frane, Andrew

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Type I Error Control in Psychology Research: Improving Understanding in General and Addressing Multiplicity in Some Specific Contexts

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Psychology

by

Andrew Vann Frane

2019

ABSTRACT OF THE DISSERTATION

Type I Error Control in Psychology Research:

Improving Understanding in General and

Addressing Multiplicity in Some Specific Contexts

by

Andrew Vann Frane

Doctor of Philosophy in Psychology

University of California, Los Angeles, 2019

Professor Martin M. Monti, Chair

The aim of this dissertation is essentially twofold: (1) to identify and correct some misunderstandings regarding Type I error control that are common in the field of psychology, and (2) to compare via simulation different multiple-testing procedures that can be used in a few specific experimental designs.

The dissertation of Andrew Vann Frane is approved.

Peter M. Bentler

Philip Kellman

Hongjing Lu

Martin M. Monti, Committee Chair

University of California, Los Angeles

2019

TABLE OF CONTENTS

LIST OF SYMBOLS AND ABBREVIATIONS

ANOVA	analysis of variance
BH	the Benjamini–Hochberg procedure
EWER	experimentwise Type I error rate
FDR	false discovery rate
FM	Félix and Menezes (2018)
FWER	familywise Type I error rate
HSD	Tukey's honestly-significant-difference procedure
k	number of simulations
m	number of hypothesis tests
m_1	number of hypothesis tests for which the null hypothesis is false
MANOVA	multivariate analysis of variance
MCP	multiple-comparisons procedure
MP	MANOVA protection
n	sample size
NSV	number of statistically significant outcome variables
PFER	per-family Type I error rate
PHEMA	the "planned-hypothesis exemption from multiplicity adjustment"
PLSD	Fisher's protected least significant difference procedure
r	sample correlation
$\widehat{\text{SE}}$	estimated standard error
v	observed incidence-rate

α	designated overall alpha level
α_C	uniform comparisonwise alpha level
α_H	nominal familywise alpha level used in the Hommel procedure
α^{**}	a Šidák-type alpha-level adjustment predicated on MANOVA significance
Δ	effect size (specifically, the standardized population mean difference)
δ^2	noncentrality parameter for the MANOVA omnibus test
ρ	population correlation
μ	population mean

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

I owe tremendous thanks to my advisor, Martin Monti, for supporting this work. I am also very grateful to the other illustrious members of my dissertation committee: Peter Bentler, Phil Kellman, and Hongjing Lu. I also wish to thank my wonderful parents—and all my labmates past and present—for assisting and encouraging me on the journey toward this dissertation's completion.

Most of the material in this dissertation either was published previously or is in-press at the time of this writing. Namely, the following papers are included as chapters in this dissertation, with minor modifications, under the same respective titles:

Frane, A. V. (in press). Misguided opposition to multiplicity adjustment remains a problem. *Journal of Modern Applied Statistical Methods*.

Frane, A. V. (2019). Some clarifications regarding power and Type I error control for pairwise comparisons of three groups. *Electronic Journal of Applied Statistical Analysis, 12*(1), 55–68. doi:10.1285/i20705948v12n1p55

Frane, A. V. (2015). Power and Type I error control for univariate comparisons in multivariate two-group designs. *Multivariate Behavioral Research, 50*(2), 233–247. doi:10.1080/00273171.2014.968836

Frane, A. V. (2015). Are per-family Type I error rates relevant in social and behavioral science? *Journal of Modern Applied Statistical Methods, 14*(1), 12–23. doi:10.22237/jmasm/1430453040

Frane, A. V. (2015). Planned hypothesis tests are not necessarily exempt from multiplicity adjustment. *Journal of Research Practice, 11*(1), Article P2. https://files.eric.ed.gov/fulltext/EJ1083896.pdf

VITA

Previously Awarded Degrees in Psychology

M.A. in Psychology California State University, Los Angeles 2014

B.A. in Psychology California State University, Los Angeles 2013

Selected Publications and Presentations

Frane, A. V. (in press). Misguided opposition to multiplicity adjustment remains a problem. *Journal of Modern Applied Statistical Methods*.

Frane, A. V. (2019). Some clarifications regarding power and Type I error control for pairwise comparisons of three groups. *Electronic Journal of Applied Statistical Analysis, 12*(1), 55–68.

Frane, A. V., & Martin, M. M. (2019). *Melody impairs reproduction of novel rhythms*. Poster presentation at the Annual Meeting of the Psychonomic Society, Montréal, Canada.

Frane, A. V. (2017). Swing rhythm in classic drum breaks from hip-hop's breakbeat canon. *Music Perception, 34*(3), 291–302.

Frane, A. V., & Shams, L. (2017). Effects of tempo, swing density, and listener's drumming experience, on swing detection thresholds for drum rhythms. *The Journal of the Acoustical Society of America, 141*(6), 4200–4208.

Frane, A. V. (2017). Errors in a program for approximating confidence intervals. *Journal of Modern Applied Statistical Methods, 16*(1), 779–782.

Frane, A. V. (2017). *Revisiting "What's wrong with Bonferroni adjustments"*. Paper presentation at the International Conference on Multiple Comparison Procedures, Riverside, CA.

Frane, A. V., & Shams, L. (2016). Clarifying some findings regarding the ventriloquist aftereffect. *Experimental Brain Research, 234*(3), 931–932.

Frane, A. V. (2016). False discovery rate control is not always a replacement for Bonferroni adjustment. *Journal of Clinical Epidemiology, 69*(1), 263.

Frane, A. V. (2016). Some clarifications regarding multiple comparisons. *Annals of Cardiac Anaesthesia, 19*(1), 144–145.

Frane, A. V. (2016). *Perceived difference between straight and swinging drum rhythms*. Poster presentation at the International Conference on Music Perception and Cognition, San Francisco, CA.

Frane, A. V. (2015). Power and Type I error control for univariate comparisons in multivariate two-group designs. *Multivariate Behavioral Research, 50*(2), 233–247.

Frane, A. V. (2015). Are per-family Type I error rates relevant in social and behavioral science? *Journal of Modern Applied Statistical Methods, 14*(1), 12–23.

Frane, A. V. (2015). Planned hypothesis tests are not necessarily exempt from multiplicity adjustment. *Journal of Research Practice, 11*(1), Article P2.

Frane, A. V. (2015). A call for considering color vision deficiency when creating graphics for psychology reports. *The Journal of General Psychology, 142*(3), 194–211.

Frane, A. V. (2015). Comment on the multiple problems of multiplicity. *The American Journal of Clinical Nutrition, 102*(6), 1619–1620.

Frane, A. V. (2015). Comment on: Epidermolysis bullosa pruriginosa: A systematic review exploring genotype–phenotype correlation. *American Journal of Clinical Dermatology, 16*(4), 335–337.

Frane, A. V. (2015). *Power and Type I error control for the Bonferroni, Hochberg, and Benjamini–Hochberg procedures*. Poster presentation at the Association for Psychological Science Convention, New York, NY.

Frane, A. V. (2015). *Color vision deficiency and graphics in psychology*. Paper presentation at the Western Psychological Association Convention, Las Vegas, NV.

Frane, A. V. (2014). *Psychology researchers disagree about multivariate analysis of variance and most of them are wrong*. Poster presentation at the American Psychological Association Conference, Washington, D.C.

Frane, A. V. (2014). *MANOVA protection vs. the Bonferroni procedure: An empirical comparison of power*. Poster presentation at the Association for Psychological Science Convention, San Francisco, CA.

Frane, A. V. (2014). *Bonferroni and beyond: An evaluation of multiple-comparisons procedures for two-group designs*. Paper presentation at the Social Science Research and Instructional Council Conference, Fullerton, CA.

GENERAL INTRODUCTION

Improving Understanding in General

Null hypothesis testing is nearly ubiquitous in psychology research (Cassidy et al., 2019; Nuijten et al., 2016). Yet among psychology researchers, there is considerable misunderstanding about Type I error control in general and about *multiplicity* (multiple hypothesis testing) in particular. For instance, an astounding number of papers in psychology and related fields have cited misguided arguments by Perneger (1998) and/or Rothman (1990) to defend unrestrained inflation of Type I error rates; for examples of such papers that were published within just a few months of this dissertation's completion, see Bekafigo et al. (2019), Bradley et al. (2019), Brinkman et al. (2019), Byrnes et al. (2019), Cameron et al. (2019), Covey et al. (2019), Dondaine et al. (2019), Dovgan and Mazurek (2019), Duffy et al. (2019), Falconer et al. (2019), Frenette et al. (2019), Groarke and Hogan (2019), Koegi (2019), Hackford et al. (2019), Jung et al. (2019), Kaseweter et al. (2019), Lange et al. (2019), Li et al. (2019), Turner et al. (2019), van der Velden et al. (2019), Van Patten et al. (2019), and others. That is not to say that the findings of those particular studies are necessarily "wrong." But I would argue that the widespread willingness to tolerate inflated rates of false findings—and to rationalize that tolerance using fallacious justifications—indicates a systemic problem in research.

To address that problem, the first three papers (chapters) that follow this introduction present ways that understanding of Type I error control can and should be improved. Specifically, the first paper debunks, and documents the impact of, "anti-adjustment" articles such as Perneger's and Rothman's; the second paper documents and critiques the widespread practice of disregarding multiplicity for "planned" comparisons, and explains in plain language

the importance of rigorous Type I error control; the third paper challenges the popular belief that the Bonferroni procedure is inherently overly conservative, and proposes that the Bonferroni procedure's strict control of the *per-family Type I error rate* may be a desirable property in many cases.

Addressing Multiplicity in Some Specific Contexts

There are often several multiple-testing procedures to choose from when planning an analysis. Therefore, in addition to encouraging thoughtful handling of multiplicity in general, this dissertation also explores which specific multiple-testing approaches are preferable in certain situations, taking both Type I and Type II error rates into consideration. Specifically, the last three papers in this dissertation consider the following experimental designs, respectively: two-group designs with multiple outcome variables, three-group designs with one outcome variable, and 2×2 factorial designs.

MISGUIDED OPPOSITION TO MULTIPLICITY ADJUSTMENT REMAINS A PROBLEM

Material in this chapter has been accepted for publication by *JMASM*.

Fallacious arguments against multiplicity adjustment have been cited with increasing frequency by researchers seeking to defend their unadjusted tests. The present paper documents the enduring impact of such arguments, and proposes that they constitute a serious problem that demands action.

Introduction

Over the many decades since Fisher (1935, pp. 64–66) informally suggested Bonferroni-type adjustment to account for multiple significance tests, a sophisticated literature has developed on how to address the problem of multiplicity. However, many studies involving multiple tests are conducted without any accounting for multiplicity whatsoever, and there is a highly influential literature advocating that practice.

Most notably, Perneger's (1998) opinion piece, "What's Wrong With Bonferroni Adjustments," which argued not only against the classical Bonferroni procedure but also against the general principle of multiplicity adjustment, has been cited over 4700 times (as per Google Scholar, June 30, 2019). A similar paper, Rothman's (1990) "No Adjustments Are Needed for Multiple Comparisons," has been cited over 3600 times (as per Google Scholar, June 30, 2019). The vast majority of studies that have cited these articles have done so uncritically and for the express purpose of defending a failure to adjust for multiplicity. The influence of anti-adjustment

articles presumably also extends far beyond the thousands of papers that have cited anti-adjustment articles directly.

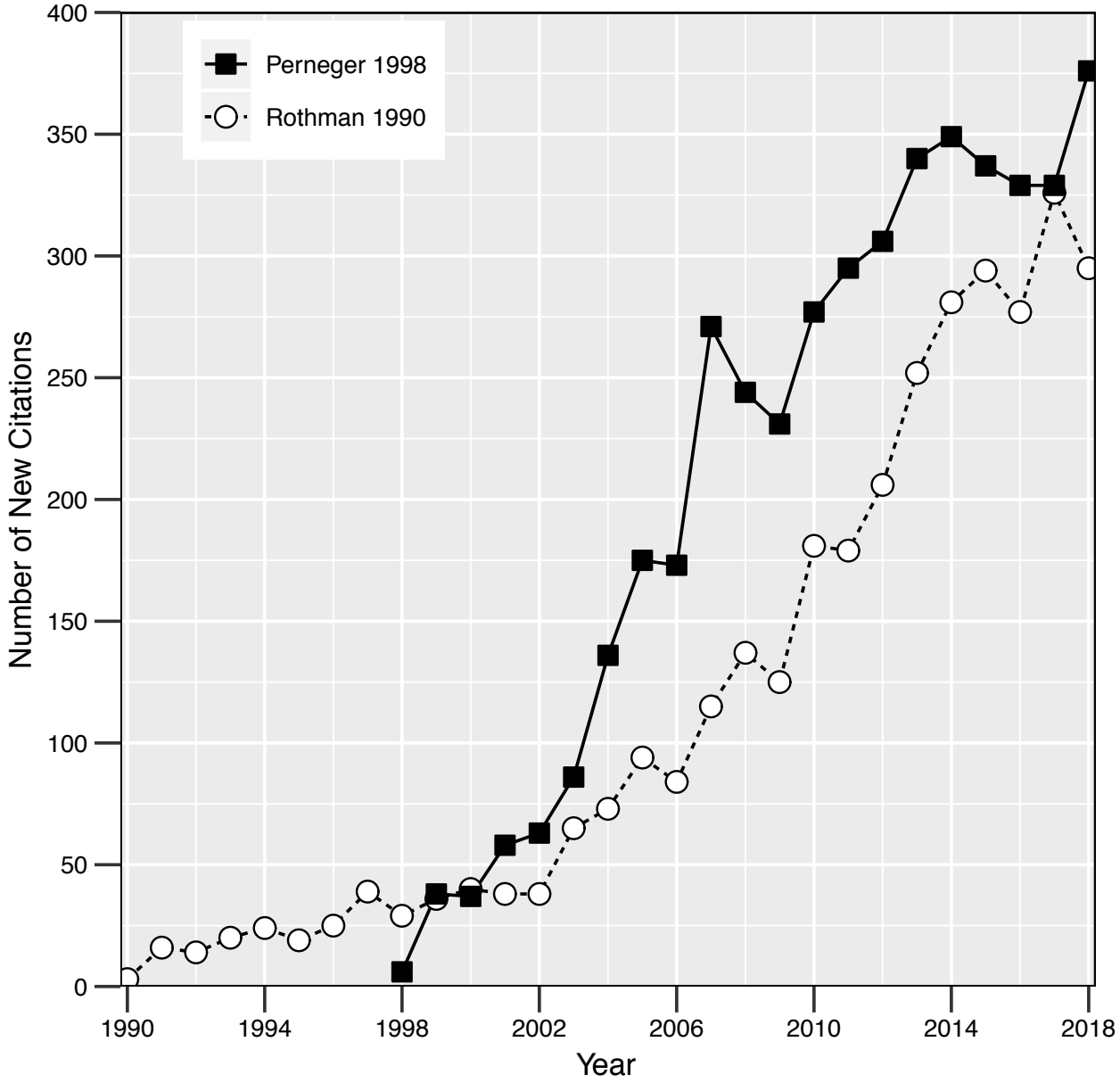

Figure 1. Number of new citations of Perneger (1998) and Rothman (1990) in each year (as per Google Scholar, April 12, 2019).

Some readers of the present paper might be inclined to dismiss anti-adjustment articles as quaint historical curiosities that could not possibly merit serious attention, especially decades after their publication. But on the contrary, papers such as Perneger's (1998) and Rothman's (1990) have maintained a remarkably enduring influence on research in a variety of scientific disciplines. In fact, the number of citations of those papers per year has trended upward over time (see Figure 1), and such citations often appear in highly regarded journals.

Moreover, the misguided arguments in classic anti-adjustment articles have reappeared in more recent papers. For instance, Mark Rubin (2017) cited both Perneger (1998) and Rothman (1990) in defense of the claim that it is fundamentally inappropriate to adjust for multiple hypotheses (see also Hurlbert & Lombardi, 2012). Some other opinion papers (e.g., Glickman et al., 2014; Nakagawa, 2007) have acknowledged the utility of controlling the *false discovery rate* in certain contexts, but have dismissed the importance of controlling the *familywise error rate* altogether—citing classic anti-adjustment papers, such as Perneger's, in support of that position. Note that false discovery rate control, though appropriate in some circumstances, is not an adequate substitute for familywise error rate control in general (Dmitrienko et al., 2010a, p. 39; Finner & Roters, 2001; Frane, 2016; Goeman & Solari, 2014; Meijer & Goeman, 2016), as the statisticians who introduced false discovery rate control were careful to emphasize (Benjamini, 2010; Benjamini & Hochberg, 1995).

Researchers are often reluctant to apply multiplicity adjustment because it reduces statistical power (all other things being equal) and thus forces them to either invest in larger samples or settle for lower power. Given the constant pressure on researchers to obtain publishable, statistically significant discoveries (Boulbes et al., 2018), perhaps it should not be surprising that anti-adjustment arguments are so popular. In some cases, self-serving researchers

may cite anti-adjustment arguments that they know are fallacious. In other cases, naive researchers may cite anti-adjustment articles in good faith, trusting in what superficially appear to be reputable sources. Indeed, in addition to a "crowd pleasing" message, anti-adjustment articles tend to have other characteristics that likely appeal to statistically unsophisticated readers who are highly vulnerable to misinformation. Namely, anti-adjustment papers typically are written in plain language, appear in non-statistical journals (with rare exceptions; Saville, 1990), and rely largely on specious appeals to "common sense" (e.g., Perneger, 1998, p. 1236) rather than on legitimate mathematical principles or on citations of statistical literature. Moreover, the recommendations in anti-adjustment articles are often simple heuristics that require little thought to implement because they advocate forgoing adjustment altogether in nearly all circumstances, with little to no consideration of contextual factors (such as the goals of the study, how the results will be used to make decisions or draw conclusions, and whether there is a hierarchical structure to the testing).

It is likely that most readers of the present paper do not need convincing that multiplicity adjustment is often important. Thus, the aim here is not to "preach to the choir" in that regard. Nor is the aim to establish procrustean rules about how multiplicity should be handled in all situations. Rather, the aim of the present paper is to document the prevalence and impact of fallacious arguments against the very principle of multiplicity adjustment, and to arm serious scientists with information that can be used to combat poor statistical practice and curb the proliferation of "statistical myths."

Before proceeding, two points should be clarified. First, the present paper often uses the term *multiplicity adjustment* in a broad sense, to include all sound methods of addressing multiplicity, even methods that do not involve adjustment per se of *p*-values or alpha levels (e.g.,

certain sequential testing methods, when the sequence is defined in a pre-registered study protocol). Second, although the present paper discusses multiplicity adjustment largely in the context of null hypothesis testing, the same principles may apply when using confidence intervals, rather than *p*-values, as the primary basis for inference or decision-making (Phillips et al., 2013). Thus, contrary to what some authors have implied (e.g., Huisingh & McGwin, 2012), examining effect-size estimates and confidence intervals rather than only *p*-values—though generally a good idea—does not in itself eliminate the problem of multiplicity. In fact, many adjustment procedures can be straightforwardly applied to confidence intervals (e.g., Dunn, 1958, 1961; Dunnett, 1955; Tukey, 1953; Westfall, 1985).

Misconceptions Underlying Anti-Adjustment Arguments

Regarding the "Universal Null Hypothesis"

Some anti-adjustment articles (e.g., Perneger, 1998; Savitz & Olshan, 1995) have incorrectly claimed that Bonferroni-type adjustments only allow inference about the "universal null hypothesis," i.e., about whether the null hypotheses are true for all tests—a view that Goeman and Solari (2014, p. 1955) rightly identified as a "myth." For example, Perneger (1998, p. 1236) claimed that if two groups are compared on 20 variables and at least one *p*-value is significant at the Bonferroni-adjusted level, "We can say that the two groups are not equal for all 20 variables, but we cannot say which, or even how many, variables differ…A clinical equivalent would be the case of a doctor who orders 20 different laboratory tests for a patient, only to be told that some are abnormal, without further detail." As is likely obvious to most readers of the present paper, that description would be true of a single omnibus test, not of multiple Bonferroni-adjusted tests. Bonferroni adjustments, and many similar methods, allow

statements to be made about each individual hypothesis because they control the familywise error rate "in the strong sense," meaning even when only some of the individual null hypotheses are true (Goeman & Solari, 2014). In fact, classical Bonferroni adjustment also controls the *per-family error rate* (i.e., the expected number of Type I errors), which is an even stricter standard than the familywise error rate (Frane, 2015a).

Like Perneger (1998), Rothman (1990) criticized multiplicity adjustment for allegedly only being relevant to the universal null hypothesis. In fact, Rothman suggested that even *entertaining* a universal null hypothesis would be fundamentally absurd: "Whereas we can imagine individual pairs of variables that may not be related to one another, no empiricist could comfortably presume that randomness underlies the variability of all observations…To entertain the universal null hypothesis is, in effect, to suspend belief in the real world and thereby to question the premises of empiricism" (pp. 44–45). However, even if only two null hypotheses are true, the familywise error rate can be inflated to approximately twice the nominal level. Thus, acknowledging Type I error inflation does not require ascribing all observable associations in the world to pure randomness. Moreover, even if all null hypotheses are false, "random" variation can still substantially affect observations, as observed associations vary in magnitude—and sometimes direction—from one sample to the next. To deny that fact would truly be to "suspend belief in the real world."

Numerous papers have uncritically cited Perneger's (1998) and Rothman's (1990) false claims about the universal null hypothesis, including papers in highly regarded journals such as *JAMA* (van Gils et al., 2009), *The Lancet* (Shulz & Grimes, 2005), and others (e.g., Armstrong, 2014; Berry, 2012; De Pablo-Fernandez et al., 2017; Glickman et al., 2014; Jenkins et al., 2009; O'Connor et al., 2009, Ostendorf et al., 2017; Racette et al., 2005; Sinclair et al., 2013; Zintzaras

& Lau, 2008). Many of those papers have in turn garnered multiple citations of their own for the same false claims. For instance, Armstrong's (2014) endorsement of Perneger's false claim about the universal null hypothesis has in turn been uncritically cited by several other authors (e.g., Day & Thorn, 2017; Kim et al., 2015; Ozcan et al., 2017; Tang et al., 2019) to defend their own unadjusted tests, demonstrating how infectious a statistical myth can be when it tells researchers what they want to hear.

The false claims about the universal null hypothesis by Perneger (1998) and Rothman (1990) have also been uncritically cited in textbooks (e.g., Ahlbom, 1993, p. 52; Aschengrau & Seage, 2014, pp. 322–323; Shulz & Grimes, 2006, p. 192). Additionally, an education research group at Stanford University responded to criticism of their unadjusted testing by claiming that adjustment is unnecessary when the universal null hypothesis is not of interest (CREDO, n.d.). Notably, that response cited only two sources: Perneger (1998) and Rothman (1990).

Regarding the Inherent Implausibility of Chance Associations

Many of Rothman's (1990) objections to multiplicity adjustment apparently reflect a more general objection to Type I error control and to any concern that observed associations in a sample might arise by chance. In Rothman's view, "Being impressed by an extreme result should not be considered a mistake in a universe brimming with interrelated phenomena" (p. 46). It is true that real associations are plentiful in the universe, but it is also true that finite samples can contain misleading associations that do not accurately reflect real effects in the population those samples are intended to represent. If that were not the case, then there would be no need for inferential statistics at all—even in the absence of multiplicity. Yet Rothman implied that misleading associations are inherently unlikely, at least in biological data.

Rothman (2014, p. 1063) doubled down on that view in a more recent article, which

contained the following non sequitur: “If one is studying experiments on psychic phenomena, skepticism about the results might lend support to multiplicity adjustments. If one is studying physiologic effects of pharmaceutical agents, real associations are to be expected and the adjustments are more difficult to defend.” But multiplicity adjustment is a mathematical correction based on the number of associations examined, not an expression of general skepticism about a given association. After all, a single positive test of “psychic phenomena” would presumably merit considerable skepticism, even though there would be no multiplicity to adjust for. Moreover, disregarding multiplicity when evaluating the efficacy of pharmaceutical products would be in direct opposition to the guidelines of regulatory agencies (European Agency for the Evaluation of Medicinal Products, 2002; U.S. Department of Health and Human Services, 1998).

To be clear, there are indeed situations in which it is appropriate to incorporate prior probabilities into the analysis. But that is not achieved by simply ignoring multiplicity. One might persuasively argue that “strictly true” null hypotheses (meaning there is no effect even negligibly different from zero) are in fact rare in biological contexts, and that researchers should therefore focus on effect size estimation rather than on null hypothesis testing. But certainly one cannot persuasively argue in favor of unadjusted null-hypothesis testing while simultaneously arguing against the relevance of null-hypothesis testing altogether. Moreover, even when focusing on effect sizes rather than on *p*-values, multiplicity adjustment can be useful in computing simultaneous confidence intervals for those effect sizes.

Regarding Statistical Power and Type II errors

Given that the entire purpose of null hypothesis testing is to protect against spurious discoveries, it would be nonsensical to defend the use of an arbitrarily high alpha level simply by

noting that high alpha levels make discoveries easier to claim. Yet many authors have effectively done just that, in defending their arbitrarily inflated familywise alpha levels by noting that unadjusted tests provide more statistical power and lower chance of Type II error. Indeed, the scientific literature contains numerous statements like the following, by Fekkes et al. (2006, p. 1570): "No adjustment for multiple comparisons, such as the Bonferroni correction, was done, because this would result in an increase in Type II errors, that is, finding a true difference and not considering this significant (Perneger, 1998)" (citation in original). Roberts et al. (2011, p. 1558) offered a similar defense of their unadjusted testing: "To avoid Type II errors no adjustment was made for multiple comparisons (Perneger, 1998)" (citation in original).

At least nine papers (Berk et al., 2014, 2017; Carral–Fernández et al., 2016; Cotton et al., 2010; Cotton et al., 2013; González–Blanch et al., 2015; Marion–Veyron et al., 2015; Mossaheb et al., 2013; Rajapakse et al., 2014) have included the following sentence, word-for-word: "No adjustments were made for multiple comparisons because they can result in a higher type II rate [sic], reduced power, and increased likelihood of missing important findings (Rothman, 1990)" (citation in originals). Nearly identical sentences have appeared in other papers (e.g., Allott et al., 2015; Smyth et al., 2015), as well as in authors' responses to reviewers' criticisms (Springer, 2016; for the corresponding published paper, see Jacka et al., 2017).

Perneger (1998) proposed several scenarios in which multiplicity adjustment would allegedly cause catastrophic Type II errors. Some of those scenarios were patently nonsensical and bore no resemblance to contexts in which multiplicity adjustment would—or perhaps even could—actually be applied, e.g., "In a clinical setting, a patient's packed cell volume might be abnormally low, except if the doctor also ordered a platelet count, in which case it could be deemed normal" (p. 1236). Some other scenarios Perneger proposed were more vaguely defined.

For example, he warned that by applying multiplicity adjustment, "an effective treatment may be deemed no better than placebo" (p. 1236). It is not clear exactly how Perneger imagined that would happen, because he did not define what the multiple tests would be in that scenario, or how the testing would be structured. In many cases, testing can be structured so that the familywise error rate is controlled without sacrificing statistical power in the primary test of treatment efficacy (European Agency for the Evaluation of Medicinal Products, 2002). And of course, if unanimous statistical significance were required on all outcomes simultaneously for the treatment to be approved, then there would effectively be no Type I error inflation to adjust for. In some other cases, multiplicity adjustment would be required—and for good reason. For instance, if a treatment were compared to placebo on five outcomes, any one of which on its own could earn approval for the treatment, then without adjustment the probability of erroneously declaring the treatment effective would be approximately 23% (given a nominal alpha level of .05, true null hypotheses, roughly independent outcomes, and satisfaction of assumptions).

Clearly, noting that looser Type I error control can provide greater statistical power is a trivial and unpersuasive argument for sacrificing statistical rigor. Although statistical power is certainly important, the proper way to limit Type II errors is by using an adequate sample size—not by allowing Type I errors to be arbitrarily inflated (European Agency for the Evaluation of Medicinal Products, 2002; U.S. Department of Health and Human Services, 1998).

In some early-stage research, it may not be feasible to collect a sample large enough to provide ample statistical power while stringently controlling for multiplicity. But in such cases, rather than ignoring multiplicity to make observed trends appear "significant," it would be more appropriate to refrain from making inferential claims until the trends are confirmed in a legitimately higher-powered study. Indeed, despite what has been suggested in some textbooks

(e.g., Aschengrau & Seage, 2014, p. 323; Savitz, 2003, p. 249), statistical nonsignificance does not necessarily imply that the null hypothesis must be "accepted" per se (in the epistemic sense) without any further investigation. Rather, statistical nonsignificance means that the null hypothesis cannot be rejected on the basis of the present evidence. Yet Rothman (1990, p. 46) claimed that multiplicity adjustment "shields some observed associations from more intensive scrutiny by labeling them as chance findings." Although that claim may accurately depict how some researchers misinterpret or misuse statistical nonsignificance in some cases—whether multiplicity is present or not—it does not constitute a legitimate criticism of the principle of multiplicity adjustment.

Regarding "Arbitrarily" Defined Families

A popular anti-adjustment argument that resembles the fallacy of slippery slope is as follows: The number of tests to adjust for is arbitrary because that *family* of tests could theoretically be extended to include all the tests conducted in a given researcher's career or all the tests reported in a given journal (e.g., Feise, 2002; Moran, 2003; Perneger, 1998; Rubin, 2017; Savitz, 2003, pp. 252–253; for similar arguments, see Huisingh & McGwin, 2012; Rothman, 1990). Considering all the tests conducted in an investigator's career or in the history of a journal would indeed be extreme ways to define the family in most cases, and the latter would present the challenge of accounting for publication bias. But considering each test in isolation would be an extreme approach in its own right. For typical applications, a middle ground is likely the most sensible strategy (Miller, 1981, pp. 31–32). After all, the typical consumer of a study report that contains multiple tests is presumably interested in the results of a particular investigation—not in the results of the author's entire career or of the journal's entire history. That said, if in a particular case there were some compelling reason to interpret results in

the context of a researcher's entire career, then it could in fact make sense to adjust inference accordingly. Notwithstanding situations where the definition of the family is dictated by some regulatory agency or other authority, "There are no hard-and-fast rules for where the family lines should be drawn, and the statistician must rely on his [or her] own judgment for the problem at hand" (Miller, 1981, p. 35).

Thus, the grouping of tests into families is contextually dependent and somewhat subjective, but not completely arbitrary. Note that the same description—"somewhat subjective, but not completely arbitrary"—could just as easily apply to numerous other *a priori* decisions, such as what sample size is sufficient, what minimum effect size to consider clinically significant, and what overall alpha level (.05 or some other level) is appropriate. Just as those decisions can be made in a thoughtful, principled way, so can decisions regarding the definition of the family. Contrary to Perneger's (1998, p. 1236) claim that "Most proponents of the Bonferroni method would count at least all the statistical tests in a given report as a basis for adjusting P values," it is doubtful that any competent statistician would recommend, for example, adjusting the confirmatory test of primary interest to account for a set of descriptive follow-up tests (European Agency for the Evaluation of Medicinal Products, 2002). In short, how the family should be defined may be debatable in some cases, but that does not mean that any definition of the family is as good as another.

Regarding Planned Tests

It is often said that hypothesis tests that have been planned *a priori* do not require multiplicity adjustment. Indeed, statements such as the following, by Fish et al. (2007, p. 1325), are common in the scientific literature: "Whilst it is true that if the Bonferroni adjustment was applied in the following analysis, none of the associations would reach the corrected threshold,

there are views strongly opposing the use of such corrections in analyses where *a priori* hypotheses exist (Perneger 1998)" (citation in original). Moreover, many textbooks on applied statistics have explicitly recommended not adjusting for multiplicity if the tests were planned (e.g., Ha & Ha, 2012, p. 206; McKillup, 2012, p. 163; Pagano, 2013, p. 422; Rutherford, 2011, p. 76; Scheff, 2016, p. 112). But there is no apparent scientific basis for that recommendation. For an exhaustive critique of the "planned-hypotheses exemption from multiplicity adjustment," see Frane (2015b).

Note also that if no specific tests are planned, then the number of potential tests for the researcher to choose from may be indeterminate, making meaningful adjustment impossible (Hochberg & Tamhane, 1987, p. 10). In that situation, researchers should not have a false sense of security that they can prevent Type I error inflation by adjusting merely for the tests that were formally conducted.

Conclusions

Although anti-adjustment arguments are frequently cited in scientific literature, they are based largely on misconceptions and, perhaps in some cases, on willful misrepresentations. Researchers should be wary of citing a mere opinion as justification for a particular approach, even if—or perhaps especially if—that opinion tells them what they want to hear.

Educators and textbook authors should warn students about common misconceptions regarding multiplicity. And they should inform students about alternatives to the Bonferroni procedure that are not as restrictive. Additionally, reviewers and editors should be aware that misconceptions about multiplicity are prevalent in the literature, and should combat the propagation of those misconceptions whenever possible. For instance, when reviewing a

manuscript, they should be on the lookout for uncritical citations of certain papers (e.g., Perneger, 1998; Rothman, 1990) that have become go-to references for researchers seeking to shield their unadjusted testing from criticism.

Once a paper that endorses a fallacious anti-adjustment argument has been published, other researchers can write critical letters in response. However, such letters typically receive much less attention than the offending article itself. For example, a letter by Aickin (1999) correctly noted that Perneger's (1998) paper "consists almost entirely of errors," and a letter by Bender and Lange (1998) was similarly critical of Perneger's paper—yet those letters could not stop the growing influence of Perneger's paper over the ensuing decades (as evident from Figure 1 in the present paper). Note that those letters did not merely present opinions, but also identified objective factual errors, most notably Perneger's claim that the Bonferroni procedure only addresses the universal null hypothesis. Yet the journal never issued any corrections to Perneger's paper.

There is widespread concern in the sciences (e.g., Baker, 2016; Open Science Collaboration, 2015; but see Jamieson, 2018) that too many findings are not replicable and that there is a high prevalence of Type I errors in the literature. Naturally, neglecting multiplicity exacerbates those problems (as noted by Bretz & Westfall, 2014; Forstmeier et al., 2016; Young, 2008). Therefore, in the present author's opinion, researchers and statisticians have a scientific responsibility to directly confront bad practice and misguided thinking concerning multiplicity. Self-serving citations of opinions and myths in order to excuse a broad disregard for statistical rigor should no longer be tolerated.

PLANNED HYPOTHESIS TESTS ARE NOT NECESSARILY EXEMPT FROM MULTIPLICITY ADJUSTMENT

Scientific research often involves testing more than one hypothesis at a time, which can inflate the probability that a Type I error (false discovery) will occur. To prevent this Type I error inflation, adjustments can be made to the testing procedure that compensate for the number of tests. Yet many researchers believe that such adjustments are inherently unnecessary if the tests were "planned" (i.e., if the hypotheses were specified before the study began). This longstanding misconception continues to be perpetuated in textbooks and continues to be cited in journal articles to justify disregard for Type I error inflation. In this paper, I critically evaluate that myth and examine its rationales and variations. To emphasize the myth's prevalence and relevance in current research practice, I provide examples from popular textbooks and from recent literature. I also make recommendations for improving research practice and pedagogy regarding this problem and regarding multiple testing in general.

Background

Null Hypothesis Testing

The *null hypothesis* is the hypothesis that a particular independent/grouping variable has no effect on (or no association with) a particular outcome variable. Often, the null hypothesis is the hypothesis that the researcher's prediction is wrong. For instance, if a researcher predicts that a particular treatment reduces depression in humans (on average), then the null hypothesis is that the treatment does not work. If a researcher predicts that a certain genetic allele is associated with Alzheimer's disease, then the null hypothesis is that the allele has no association with

Alzheimer's disease. However, the null hypothesis applies even when the researcher makes no official prediction, so long as there is a possibility that there is no effect/association.

Because hypotheses typically cannot be tested on the entire population of interest (e.g., by analyzing the genomes of every living human being), hypotheses are instead tested on a finite sample of the population. Thus, a researcher never knows with 100% certainty whether an ostensible effect that is observed in the sample actually exists in the population or whether it is merely due to "chance." For instance, despite random assignment, a treatment group may happen to be, on average, more predisposed to improve than the subjects in a placebo group.

In conventional (frequentist) hypothesis testing, the researcher addresses this inevitable uncertainty by computing a *p*-value based on the observed data. Roughly speaking, the *p*-value represents the theoretical probability that the observed effect (or a larger effect) would occur by chance if the null hypothesis were true. Once computed, the *p*-value is then compared to a predesignated critical value called the *alpha level* (α), such that if $p < \alpha$, then the null hypothesis may be rejected. Once the null hypothesis is rejected, the observed effect may be declared *statistically significant*, and a corresponding decision can be made (e.g., a treatment is recommended, an association is claimed, a follow-up study is pursued, etc.).

A statistically significant result that occurs when the null hypothesis is true is called a *Type I error*. Hence, α represents the maximum *Type I error rate* that the researcher is willing to tolerate. For example, among tests that use the conventional .05 alpha level, a Type I error is allowed to occur up to 5% of the time.

Type I error rates can be reduced by making alpha levels lower (i.e., more stringent), but only at the expense of *statistical power* (the likelihood of producing statistically significant results when the null hypothesis is false). Because the goal of research is frequently to

discover/demonstrate some effect or association, and because researchers typically face considerable pressure to find statistical significance (e.g., in order to get published or promoted), researchers are often reluctant to sacrifice statistical power.

Another way to reduce the effective Type I error rate is to require that significant results be promptly replicated by a second study with a completely new sample. In terms of the effective Type I error rate, making statistical significance conditional on two independent tests, each at α, is equivalent to conducting a single test at α^2 (e.g., at .0025 when nominal $\alpha = .05$). However, immediate full-scale replications are rare, largely for practical reasons. More commonly, significant results are reported shortly after they are obtained, rather than withheld pending an independent corroboration.

The Problem of Multiple Testing

The Type I error rate is fairly straightforward when there is only one test. However, scientific research often involves testing more than one hypothesis at a time, e.g., when evaluating more than one mean difference or more than one correlation. The resulting problem of *multiplicity* (multiple testing) is well known: Every hypothesis test added to a data analysis carries additional potential for error, so the *testwise alpha levels* (i.e., the nominal alpha levels at which the individual tests are conducted) can substantially understate the effective Type I error rate for the investigation as a whole. For example, when two tests are conducted, each at the .05 level, the probability that at least one of them would produce a Type I error if both hypotheses were true may be as high as .10, though the exact probability depends on the statistical *dependence* (correlation, in the general sense) between the tests.

Thus, if Type I errors are to be *controlled* (i.e., contained at a given rate), then adjustments should be made to compensate for the number of tests in the *family* (the set of tests

being examined). These adjustments, sometimes called "corrections," typically involve reducing testwise alpha levels (or equivalently, adjusting *p*-values upwards), thereby reducing statistical power. However, multiplicity adjustments also apply to the widths of *confidence intervals*, even when *p*-values are not used (Benjamini & Yekutieli, 2005; Dunn, 1961; Hsu, 1996; Miller, 1981). Confidence intervals are computationally related to null hypothesis tests, but are used to make inferences about the effect sizes, rather than merely about whether the effects are zero or nonzero. Note that although this paper generally discusses multiplicity in terms of null hypothesis testing, the same principles of multiplicity may be relevant to computing confidence intervals.

Ways to Define the Type I Error Rate in Multiple Testing

Many *multiple testing procedures* (i.e., methods of adjustment for multiplicity) have been devised. Which multiple testing procedure is preferable for which situation is a complex question that cannot be definitively answered, but using no method at all is clearly a poor default strategy. In any case, before choosing a multiple testing procedure, one should first decide which error rate is relevant for the given investigation (Benjamini, 2010). Many error rates have been defined, most notably the following three, in order of decreasing stringency; note that each of these three error rates is equal to the testwise alpha level when there is only one test, but inflates as the number of tests increases.

Per-Family Type I Error Rate (PFER; Tukey, 1953). The PFER is the expected number of Type I errors per family. Note that the "expected number" is a long-term average, not an upper bound on the number of Type I errors likely to occur in any single investigation. The PFER is typically controlled using the Bonferroni procedure, which can be applied to any set of *p*-values by setting the testwise alpha level at α / m, where α is the designated *overall alpha level*

and m is the number of tests. The Bonferroni procedure can be similarly applied to confidence intervals, by expanding the width of each interval at the nominal $1 - \alpha$ confidence level to what it would be at the $1 - \alpha / m$ confidence level (Dunn, 1961).

Familywise Type I Error Rate (FWER; Tukey, 1953). The FWER is the probability that at least one Type I error will occur in a given family. Thus, FWER control is more permissive of Type I errors than PFER control is, because multiple simultaneous errors do not add to the tally of "at least one Type I error" any more than a single error does. However, in many cases, the FWER is only negligibly lower than the PFER, especially when the number of tests is small and the dependency among the tests is low (because simultaneous Type I errors are relatively rare under such conditions).

The Bonferroni procedure is often described as controlling the FWER, which it does, because any procedure that controls the PFER at α controls the FWER at $\leq \alpha$. However, by sacrificing strict PFER control, other FWER-controlling procedures (e.g., Holm, 1979; Hommel, 1988) can provide more statistical power; see Dmitrienko et al. (2010a) for a litany of such procedures, each with its own advantages and limitations. Thus, given the multitude of FWER-controlling procedures available, the oft-lamented "conservatism" of the Bonferroni procedure is not an adequate excuse for forgoing FWER control altogether.

It is important to distinguish FWER control from "weak FWER control," which is FWER control that is reliable when all null hypotheses are true, but can fail when one or more null hypotheses are false. Weak FWER control is typically achieved by making several simultaneous tests (none of which are adjusted) conditional on the statistical significance of a single omnibus test (e.g., ANOVA or MANOVA), a technique that is sometimes called "protected" testing. Because this approach does not reliably control Type I error (except in certain circumstances;

Bird & Hadzi–Pavlovic, 2014; Levin et al., 1994), it has very limited applicability (Benjamini, 2010; Goeman & Solari, 2014; Hsu, 1996; Tamhane, 2009). In fact, most methods of Type I error control (e.g., see Dmitrienko et al., 2010a) do not require omnibus tests at all.

False Discovery Rate (FDR; Benjamini & Hochberg, 1995). The term *false discovery* is generally synonymous with *Type I error*, but the term *FDR* refers to one particular form of Type I error rate. Loosely speaking, the FDR is the expected proportion of statistically significant tests that are Type I errors in a given family (except when all null hypotheses are true, in which case the FDR is equivalent to the FWER). Note that the expected proportion is a long-term average, not an upper bound on the proportion of statistically significant tests likely to be false in any single investigation. Note also that the computation of this long-term average defines the proportion as zero when no tests are significant.

Any procedure that controls the FWER at α controls the FDR at $\leq \alpha$, but by sacrificing strong FWER control, dedicated FDR-controlling procedures can provide more statistical power. FDR control can be useful when there are numerous tests and allowing some Type I errors is not very harmful (e.g., when screening for associations to be examined in subsequent studies). However, FDR control is not sufficient when stronger, more confirmatory inference is required (Benjamini, 2010; Dmitrienko et al., 2010a; Goeman & Solari, 2014; Meijer & Goeman, 2016). Note also that the FDR's relevance is limited when hypotheses have unequal likelihoods, because tests that are known to produce low *p*-values (tests that could be called "ringers") can drive down the FDR, thereby allowing tests with higher *p*-values to become statistically significant (Finner & Roters, 2001).

Scientific Harm Caused By Type I Errors

Subjecting hypotheses to rigorous testing is a cornerstone of the scientific method. If

false discoveries were inconsequential, then researchers' speculations and intuitions could simply be declared correct without being tested at all. However, false discoveries can cause "scientific harm," e.g., by impeding scientific progress, misdirecting scientific understanding, impairing scientific credibility through poor *replicability* (reproducibility of results), and causing resources to be squandered on spurious findings. Hence, although Type I errors cannot be eliminated, they should be controlled.

Of course, "missed true discoveries" (*Type II errors*) can be scientifically harmful in their own way, which is why it is important to use sample sizes that provide adequate statistical power. However, Type II errors are arguably more likely to be corrected than Type I errors in many cases, because they tend to be less reinforced by factors such as confirmation bias and publication bias, and because promising leads are unlikely to be abandoned without a second look simply because statistical significance was missed by some nominal amount; note that a *failure to reject* the null hypothesis does not necessarily constitute an *acceptance* of the null hypothesis. Moreover, as Ryan (1962) opined regarding the comparative threats of Type I and Type II errors in psychology research, "I believe that it is less important if we miss some very small effect of a variable, than it is to claim that the variable has an effect (of unspecified magnitude) which does not actually exist at all." Note also that uncontrolled Type I error rates threaten the credibility even of true discoveries, as statistical significance ceases to be meaningful when it is too easily achieved by chance.

By limiting the rate at which false discoveries are allowed to occur, hypothesis testing provides some protection against the scientific harm caused by false discoveries. The purpose of multiplicity adjustment is simply to preserve that limit when there are multiple simultaneous opportunities for harm. Hence, multiplicity adjustments should account for each *opportunity for*

harm, i.e., each test that would constitute a discovery on its own if statistically significant. The number of potential discoveries in a given study often is straightforward, but other times is subjective. As the following two examples illustrate, whether certain tests qualify as potential discoveries depends on how the results might be used:

> First, consider a 2 (teaching method: old, new) × 2 (student's sex: male, female) factorial design with three planned orthogonal contrasts: main effect for teaching method, main effect for gender, and an interaction, with some measure of student achievement as the dependent variable. Imagine that the researchers will publish their findings if any of the three contrasts are statistically significant. In this case, the probability of publishing a false discovery can be nearly three times the testwise alpha level, so adjustment for multiplicity is likely advisable.
>
> On the other hand, imagine that for the same 2 × 2 design and the same three contrasts, the goal of the study is to get approval to replace the old teaching method with the new one, i.e., the goal is to demonstrate a main effect for teaching method. Imagine that the other contrasts are merely descriptive (e.g., to verify an assumption that student's sex is irrelevant to achievement in the course). Multiplicity adjustment is arguably not necessary in this case, because the opportunity for a harmful false discovery is confined to a single contrast: main effect for teaching method. A main effect of student's sex could make an interesting refinement of the results, and a method-sex interaction could be a relevant caveat to the results, but only a main effect of teaching method has the potential to generate approval for the new method.

Table 1

Selected Studies From 2014 That Defended Their Unadjusted Multiple Testing by Noting That the Tests Were Planned A Priori

Study	Journal	Excerpt
Cachelin et al. (2014)	*Cultural Diversity and Ethnic Minority Psychology*	"The t-tests were planned and hypothesis driven, therefore no adjustment for multiple testing was employed."
Fenesi et al. (2014)	*The Journal of Experimental Education*	"All post hoc *t* tests were Bonferroni corrected to *p* [sic] < .05; a priori planned comparisons were not (Perenger [sic], 1998; Rothman, 1990)."
Glaus et al. (2014)	*Journal of Psychiatric Research*	"P-values were not adjusted for multiple testing because the hypothesized associations between mental disorders and inflammatory markers were specified a priori."
Holmes et al. (2014)	*Mutation Research: Fundamental and Molecular Mechanisms of Mutagenesis*	"Since all comparisons among means were considered to be of substantive interest a priori, no adjustment for multiple comparisons was incorporated into the analysis."
Krane–Gartiser et al. (2014)	*PLoS ONE*	"A correction for multiple comparisons adjusting for the total number of statistical tests has not been done since the analyses were planned before they were conducted."
MacDonald & Barry (2014)	*International Journal of Psychophysiology*	"Since all contrasts were planned and there were no more of them than the degrees of freedom for effect, no Bonferroni-type adjustment to α was necessary."
Pataki et al. (2014)	*Journal of Early Childhood Literacy*	"No correction for multiplicity was employed as our *a priori* intent was to test each variable independently."
Pyra et al. (2014)	*Journal of General Internal Medicine*	"All analyses were planned a priori; therefore, p values were not adjusted for multiple comparisons."
Stenfors et al. (2014)	*BMC Psychology*	"Since the significance tests were used to evaluate a set of a priori hypotheses, individual test results were not corrected for multiple significance testing."

Clearly, the harm caused by Type I and Type II errors must be evaluated on a case-by-case basis. There are other subjectivities to consider as well. For example, researchers

may disagree on whether a particular study containing three experiments should be considered to have three distinct families of hypotheses, or whether all the tests in the study should be considered as a single family and adjusted accordingly. And even in the absence of multiplicity, researchers may disagree on what overall alpha level is appropriate, as there is no particular scientific specialness to the .05 level and some questions presumably require more confident answers than others.

However, the fact that there is subjectivity regarding an issue does not mean that all statements about that issue are equally valid. For example, it would not be sensible to say, "Because there is subjectivity regarding what alpha level is appropriate, it is therefore appropriate to test all my hypotheses at $\alpha = .99$." Nor is it sensible to say, "Because there is subjectivity regarding how multiplicity should be handled, it is therefore appropriate to disregard multiplicity." On the contrary, subjective issues frequently require more thoughtful consideration than objective issues.

The Planned-Hypotheses Exemption From Multiplicity Adjustment

As numerous authors have noted (e.g., Anderson, 2014; Glickman et al., 2014; Ha & Ha, 2012; Iacobucci, 2001; O'Keefe, 2003; Rutherford, 2011; Ryan, 1959, 1995; Sheskin, 2011; Stangor, 2015; Stanley, 1957; Steinfatt, 2006; Streiner, 2015; Thompson, 1994; Tucker, 1991; Weiss, 2006), many in the applied sciences consider it appropriate not to adjust for multiplicity if the tests were *planned* (i.e., if the hypotheses were specified *a priori*, meaning before the study began). In fact, researchers have frequently defended their unadjusted tests explicitly on the basis that the tests were planned (see Table 1 for a few examples). The belief that stating one's hypotheses *a priori* eliminates or excuses Type I error inflation—a belief this paper refers to as

the *planned-hypotheses exemption from multiplicity adjustment* (PHEMA)—has no apparent mathematical or scientific basis. Yet the myth continues to be perpetuated. For example, consider the following passage from a popular textbook: "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" (Pagano, 2013, p. 422; emphasis in original). See Tucker (1991) and Wang (1993) for similar statements. Note that although PHEMA does not come with a scientific justification, it does come with a seductive offer: more statistical power.

Possible Origins of PHEMA

The term *planned comparisons* is often used in the context of ANOVA-based analyses, but more generally can refer to any tests of hypotheses (sometimes called *specific hypotheses*) that were generated *a priori*. Planned comparisons are distinguished from *unplanned comparisons*, which are performed without any a priori expectation, e.g., when relationships that were not previously considered interesting are detected in the data. Note that the number of unplanned comparisons implicitly includes not only those that are reported, but also any comparison that would have been reported had it been statistically significant (Tamhane, 2009). Consequently, if a researcher is willing to tout the relevance of any relationship that happens to turn up, then the opportunity for Type I error is inflated by every spurious relationship that could potentially appear. Thus, it is true that controlling Type I error for all conceivable tests (e.g., *all possible comparisons*) typically requires more severe adjustment (and hence "costs" more in statistical power) than controlling Type I error for only a predetermined subset of tests (Cohen et

al., 2003; Hsu, 1996). But unfortunately, that truth seems to have been distorted into the myth that planned comparisons do not require adjustment at all.

Ryan (1995) blamed this confusion partly on ambiguous use of the term *post hoc*, which means "formulated after the fact." For example, the phrase *post hoc tests* is often used to mean unplanned tests (i.e., tests conceived *post* data-collection), but is sometimes used to mean multiple tests in general, especially multiple tests that follow an omnibus test. This equivocation may lead some to believe that multiple testing is only of concern for unplanned tests—a confusion that is perhaps reinforced by statistical software, such as SPSS, that list all multiplicity adjustments, including the Bonferroni procedure, as "post hoc" options (Howell, 2013).

Rationalizations for PHEMA

Greater Importance of Planned Tests

Keppel and Zedeck (1989, p. 172) noted that PHEMA "is generally defended by the argument that planned comparisons typically constitute the primary purpose of a study, and as such, they should be subjected to the most sensitive statistical test possible." However, this approach allows the most important questions (i.e., "the primary purpose" of the study) to be investigated with the least rigor (i.e., with minimal control of Type I error). Moreover, using "the most sensitive statistical test possible" only makes sense under the constraint that Type I error is controlled. Otherwise, why not set the alpha level at .99 rather than at .05? After all, if Type I error control is not of concern, then any test can be made more "sensitive" (i.e., more statistically powerful) simply by raising the alpha level. A better way to achieve adequate statistical power would be to invest in a larger sample size.

Incidentally, if a study involves one planned test of primary importance and multiple tests

of somewhat lesser interest, there is a simple way to control the FWER without reducing the sensitivity of the primary test:

Step 1: Conduct the primary test at the unadjusted alpha level.

Step 2: If the primary test is significant, then conduct the secondary tests using testwise alpha levels adjusted for the number of secondary tests. But if the primary test is not significant, then forfeit the significance of the secondary tests. Note that when using this method, the testing order and conditionality should be explicitly outlined *a priori* in a registered study protocol.

Greater Credibility of Planned Tests

Another common rationale for PHEMA is that *a priori* predictions are presumably logical extensions of extant knowledge and are therefore more likely to be correct (Abelson, 1995; Anderson, 2014; Ha & Ha, 2012; McHugh & Ellis, 1957; Rutherford, 2011). One textbook advised the following: "Because you have preplanned these comparisons, typically based on prior data and theory, and you do not plan to do *all possible* comparisons, you are not required to make a correction for your alpha (α) level" (Ha & Ha, 2012, p. 206; emphasis in original). However, that appears to be a non sequitur. It may be true that a group of predictions are generally more likely to be correct if they have some theoretical basis, but the same would be true of a single prediction. Thus, why should "preplanning" excuse relaxed Type I error control for multiple tests if preplanning would not excuse relaxed Type I error control for one test?

Dissemination of PHEMA: An Example

Even a patently false heuristic such as PHEMA can become popular if it tells people what

they want to hear, e.g., that multiple tests may be conducted without sacrificing statistical power. For instance, Perneger's (1998) manifesto against multiplicity adjustments, which promoted PHEMA and numerous other misunderstandings (as noted by Aickin, 1999; Bender & Lange, 1998; Goeman & Solari, 2014), has been cited by over 3,000 articles as of this writing—and the majority of those articles were published in 2010 or later (see http://scholar.google.com).

One such article (Roche & Chainay, 2013) defended its unadjusted tests as follows: "Because we were testing specific hypotheses, we performed planned comparisons, which, unlike post hoc tests, do not need to be adjusted. In light of criticism in the literature aimed at Bonferroni and other corrections (e.g., Perneger, 1998), the analyses were performed without adjustment." Sijbrandij et al. (2013) offered a similar justification for their unadjusted tests, also citing Perneger: "Since pre-specified hypotheses were tested, no formal corrections for multiple comparison [sic] were carried out (Perneger, 1998)." For other PHEMA-based statements citing Perneger, see Askari et al. (2013), Clifford et al. (2012), Fenesi et al. (2014), Kawai et al. (2014), Krane–Gartiser et al. (2014), Lau et al. (2012), Weisse et al. (2013), and many others.

Variations on PHEMA

Constraining PHEMA to Orthogonal Contrasts

Many textbooks have suggested that although multiplicity may be of concern for some planned tests, multiplicity is not of concern for planned orthogonal contrasts (Abdi & Williams, 2010; Brown, 1990; Cohen, 2013; Cohen et al., 2003; Doncaster & Davey, 2007; Kirk, 2013; Pedhazur & Schmelkin, 1991; Randolph & Meyers, 2013; Zieffler et al., 2011). In fact, some researchers have explicitly defended their unadjusted comparisons on that basis (e.g., Harkness & Luther, 2001; Nam & Zellner, 2011; Nieuwenhui et al., 2013).

The reasoning for this version of PHEMA may be summarized as follows (Abdi & Williams, 2010, p. 248): "Planned orthogonal contrasts are equivalent to independent questions asked to the data. Because of that independence, the current procedure is to act as if each contrast were the only contrast tested" (see also Day & Quinn, 1989; Ruxton & Beauchamp, 2008; Thompson, 1994). However, this rationale appears to depend on equivocal use of the word "independence": *Statistical independence* (i.e., mutual orthogonality) among the tests does not imply that each result should be interpreted "independently" (i.e., without regard to how many other tests were conducted).

In fact, the FWER is higher for orthogonal tests than for positively dependent tests. Specifically, the maximum FWER for unadjusted tests monotonically diminishes from $1 - (1 - \alpha)^m$ to α as the correlation among the tests increases from 0 to 1, where α is the designated alpha level and m is the number of tests. Thus, not only is adjustment for multiplicity potentially important for orthogonal contrasts (Bechofer & Dunnett, 1982), one could argue that it is *especially* important for orthogonal contrasts. Incidentally, the maximum FWER can be higher for negatively dependent tests than for orthogonal tests, but typically only marginally so, and negative dependence istypically not plausible for two-sided tests.

Constraining PHEMA to Small Numbers of Hypotheses

Another variation on PHEMA asserts that multiplicity may be disregarded for planned tests provided that the number of tests is sufficiently small. Limiting the number of unadjusted tests that may be excused by PHEMA is often recognized as necessary "because otherwise, the researcher could delineate a very long list of contrasts and claim them all as planned" (Iacobucci, 2001).

For multi-group designs, some authors have set the maximum number of unadjusted

comparisons at one less than the number of groups (e.g., Keppel & Zedeck, 1989; Tabachnick & Fidell, 2012), an approach that Ryan (1959) called "Duncan's compromise" because it was proposed by Duncan (1951, 1955). This limit is equal to the maximum number of orthogonal contrasts and also equal to the number of numerator degrees of freedom that would be available in an omnibus test. Other proposed limits on the number of unadjusted tests have been less precise, e.g., a "small number" (Armstrong, 2014, p. 505; Hays, 1988, p. 411; Helweg–Larsen & Nielsen, 2009, p. 91; McKillup, 2012, p. 163; Streiner & Norman, 2011, p. 18), or a "low" number (Baguley, 2012, p. 491), or "few" (Pagano, 2013, p. 402; Welkowitz et al., 2012, p. 364). However, all of these proposed constraints are overly permissive of Type I error inflation, given that even going from one test to two tests without adjustment can roughly double the PFER and FWER.

Moreover, allowing more Type I error inflation for a small number of tests than for a large number of tests is arbitrary and logically inconsistent. For instance, suppose that if there are only three tests, then it is deemed acceptable not to adjust for multiplicity, but that if there are ten tests, then FWER control is deemed necessary. Assuming an unadjusted alpha level of .05, the maximum FWER for three tests is roughly .14. But if .14 is an acceptable FWER for three tests, then why should .14 not be an acceptable FWER for ten tests? That is, why insist that the Type I error rate for one test should be controlled at .05, and that the FWER for ten tests should also be controlled at .05, but that the FWER for three tests may be controlled at .14?

Reverse-PHEMA

Some authors have proposed the opposite of PHEMA: that planned tests require multiplicity adjustment and that unplanned tests are exempt (e.g., Rova et al., 2014, p. 256). This

heuristic, which is no more mathematically justifiable than PHEMA, is perhaps based on an assumption that unplanned tests are typically exploratory (i.e., not confirmatory) and therefore require less rigorous control of Type I error. However, even exploratory analyses often require some form of multiplicity adjustment, as one would not want to waste resources following up on an excessive number of spurious preliminary findings (Tamhane, 2009). It is true that in some unplanned testing scenarios, the number of implicit tests may be indeterminate, making formal multiplicity adjustment impossible (Bender & Lange, 2001). However, in such contexts, *p*-values can only serve a descriptive function and should not be interpreted—or reported—as if they are hypothesis test results.

Conclusions

There is considerable concern in the sciences about poor replicability of published findings and what is perceived as a high prevalence of false discoveries (Pashler & Wagenmakers, 2012). Adequate control of Type I error inflation directly relates to those issues and is essential to good research practice and scientific soundness (Benjamini, 2010; Bretz & Westfall, 2014; Hsu, 1996). False heuristics such as PHEMA, that discourage thoughtful handling of multiplicity, are therefore a nontrivial hindrance to research quality.

That is not to say that PHEMA necessarily reflects the dominant view among researchers. For example, in confirmatory trials to demonstrate drug efficacy, comparisons are typically required to be both prespecified in the study protocol and adjusted for any multiplicity (U.S. Department of Health and Human Services, 1998; European Agency for the Evaluation of Medicinal Products, 2002). But given that so many respected textbooks have endorsed PHEMA in one form or another, and given that so many recent articles have used PHEMA to justify

forgoing multiplicity adjustment, it is evident that awareness, education, and standards of practice regarding this issue need improvement. Therefore, although the present document is not the first to criticize PHEMA (e.g., see Ryan, 1959, 1995), it aims to provide the most thorough refutation of PHEMA and its variations.

Recommendations for Researchers

• Avoid using PHEMA as an excuse for unadjusted (or under-adjusted) tests. In some cases, there may be a legitimate reason not to adjust—but PHEMA is not such a reason. Note that the mere fact that subjectivities and disagreements about multiple testing exist does not mean that the problem may be disregarded or that all statements about the problem are equally valid.

• Select an error rate appropriate for the type of inference required. For example, PFER control is appropriate when the veracity of each claimed discovery is highly important, whereas FDR control provides more statistical power and may be preferable when it is sufficient merely to have an adequate preponderance of discoveries be correct (e.g., when screening through a large number of associations to generate hypotheses for future study). In terms of stringency, FWER control occupies a middle ground between the other two rates: It considers avoiding even one Type I error important, but considers multiple simultaneous Type I errors to be no more worrisome than a single Type I error.

• As recommended by the American Psychological Association (2011) and by other sources (e.g., Tromovitch, 2012), report precise *p*-values rather than merely reporting "$p < .05$," so that readers requiring a different level of inference can apply an alternative approach. Note also that confidence intervals are generally more informative than

p-values alone, given that the size of the effect—not merely whether the effect is different from zero—is presumably important in most cases.

• Regardless of which approach to Type I error control is used, report the number of tests conducted (including those implicitly conducted when "fishing" through the data for significance), the structure of the testing (e.g., which comparisons were of primary and secondary interest *a priori*), and why the chosen approach to Type I error control was deemed appropriate for the study. An *a priori* statistical power analysis is often valuable as well, especially when nonsignificant results are potentially interesting. When possible, all this information should be preregistered in a study protocol (or similar document) before the study begins—which typically should be no problem for analyses that truly are "planned."

Recommendations for Professors and Textbook Authors

• Refrain from perpetuating PHEMA, and explicitly refute PHEMA when presenting the concept of multiplicity or when distinguishing between planned and unplanned tests.

• Be wary of the term *post hoc*, which has become ambiguous through misuse. In fact, Ryan (1995) recommended that the term not be used at all in the context of hypothesis testing. The word *exploratory* may also be problematic: The term generally means "not confirmatory," but is often used as a synonym for "unplanned" when describing a data analysis—even though planned tests can be exploratory also, especially in early stages of research.

• When discussing how statistical procedures should be applied, emphasize the fundamental goals of those procedures. For example, the purpose of null hypothesis

testing is to limit the rate at which harm is caused by false discoveries, and the purpose of multiplicity adjustments is to preserve that limit when there are multiple simultaneous opportunities for harm. If these basic goals are understood, then it is easy to recognize that whether the tests were planned or not is irrelevant to those goals—a planned opportunity is an opportunity nonetheless.

ARE PER-FAMILY TYPE I ERROR RATES RELEVANT IN SOCIAL AND BEHAVIORAL SCIENCE?

The *familywise Type I error rate* is a familiar concept in null hypothesis testing, whereas the *per-family Type I error rate* is rarely addressed. This paper uses Monte Carlo simulations, graphics, and the applied statistics literature to make a case for the relevance of the per-family Type I error rate in social and behavioral science.

Introduction

The *familywise Type I error rate* (FWER; Tukey, 1953), which is the probability of making at least one Type I error in a family of hypotheses, is a familiar concept in quantitative research. Much less frequently addressed is the *per-family Type I error rate* (PFER; Tukey, 1953), which is the number of Type I errors expected to occur in a family of hypotheses (in other words, the sum of probabilities of Type I error for all the hypotheses in the family). The unpopularity of the PFER presumably comes largely from the fact that it is a stricter standard than the FWER, meaning that controlling it can be more costly in statistical power—increasing the Type II error rate (Shaffer, 2002). Given the tremendous pressure on researchers to find statistically significant *p*-values, any reduction in statistical power is a hard sell. However, it is arguable that the PFER is often more relevant than the FWER in social and behavioral science research (Barnette & McLean, 2005; Klockars & Hancock, 1994; Ryan, 1959, 1962). The argument is essentially as follows: Committing multiple Type I errors simultaneously is worse

than committing only one, yet unlike the PFER, the FWER does not distinguish between making one Type I error in a family and making several Type I errors in a family. Moreover, one might reason that because both the maximum FWER and the maximum PFER are equal to α when there is only one comparison, both error rates should remain less than or equal to α when there are multiple comparisons if Type I error is to be considered "uninflated."

Readers may debate the comparative merits of the FWER and the PFER. The goal of this paper is not to definitively advocate for one standard over the other, but rather to point out that although both error rates have merits, the PFER is almost universally ignored and may deserve more attention. For example, in statistics textbooks for the social and behavioral sciences, there is generally no mention of the PFER even when the FWER is addressed (e.g., Goodwin, 2010; Hinton, 2004; Howell, 2014; Mertler & Vannatta, 2010; Meyers et al., 2006; Sirkin, 2006; Stevens, 2009; Tabachnick & Fidell, 2012; Wetcher–Hendricks, 2011). And although some classic texts on simultaneous inference discuss the PFER (e.g., Hochberg & Tamhane, 1987; Miller, 1966; Tukey, 1953), many newer books on the subject do not (e.g., Dickhaus, 2014; Dmitrienko et al., 2010b; Hsu, 1996).

The present study briefly describes some popular Type I error rate controlling procedures, distinguishing PFER control from FWER control. Then examples from the applied statistics literature are used to show how widespread disregard of the PFER may be causing confusion. Then Monte Carlo simulations are used to demonstrate that in multivariate contexts the PFER can be substantially inflated even when the FWER is controlled, particularly when outcome variables are positively correlated.

Controlling the PFER Using the Bonferroni Procedure

The Bonferroni procedure caps the maximum PFER at α by testing each hypothesis at a

testwise alpha level of α / m, where m is the number of hypotheses in the family. With rare exception (e.g., Harris, 2001), textbooks tend not to mention that the Bonferroni procedure controls the PFER, and instead recommend it only as a method for controlling the FWER. It is true that the Bonferroni procedure controls the FWER (as does any method that controls the PFER), but using a PFER-controlling method to control the FWER prompts two questions: (1) If the objective is to control the PFER, then why not say so, and (2) if the objective is to control the FWER, then why not use a procedure that is more optimized for that purpose? After all, several methods for controlling the FWER are more powerful (meaning they can produce significance in more comparisons) than the Bonferroni procedure. Among the most popular of these methods are stepwise procedures, such as the Holm and Hochberg procedures, which are described in the following section.

Controlling the FWER Using Stepwise Procedures

Holm's (1979) procedure first arranges the m hypotheses from lowest to highest p-value. Then the hypotheses are tested sequentially in that order, each at a testwise alpha level of $\alpha / (m - b + 1)$, where b is a number between 1 and m indicating the position of the given hypothesis in the sequence. Thus, the first hypothesis is tested at level α / m, the next at $\alpha / (m - 1)$, the next at $\alpha / (m - 2)$, and so on until the last hypothesis is tested at level α. Testing is conditional, meaning that if any p-value in the sequence is nonsignificant, then all larger p-values are also declared nonsignificant and testing stops. Holm's method controls the FWER, is more powerful than the Bonferroni procedure, and requires only slightly more computation. Unlike the Bonferroni procedure, Holm's method does not always allow computation of confidence intervals (Strassburger & Bretz, 2008; Guilbaud, 2008).

Hochberg's (1988) procedure is essentially the reverse of Holm's: The hypotheses are

arranged from highest to lowest *p*-value, and then tested sequentially in that order, each at a testwise alpha level of α / b, where b is a number between 1 and m indicating the position of the given hypothesis in the sequence. Thus, the first hypothesis is tested at level α, the second at $\alpha / 2$, the third at $\alpha / 3$, and so on until the last hypothesis is tested at level α / m. If any *p*-value in the sequence is significant, then all smaller *p*-values are also declared significant and testing stops. Hochberg's procedure controls the FWER (except in certain situations; see Dmitrienko et al., 2010a) and is more powerful than Holm's, but generally does not allow computation of confidence intervals (Dmitrienko et al., 2010a; Guilbaud, 2012).

Some other stepwise procedures for controlling the FWER are more powerful than Hochberg's (e.g., Hommel, 1988; Rom, 1990), but they are more computationally complex and, like Hochberg's method, generally do not allow computation of confidence intervals (Dmitrienko, 2010a; Guilbaud, 2012). There are also methods that control the FWER in specific contexts. For example, Dunnett's (1955) procedure and its variations (see Dmitrienko et al., 2010a) can be used when comparing multiple treatment groups to a single control group (and not to each other). There are also Šidák-based methods (see Bird & Hadzi–Pavlovic, 2013), which are not necessarily applicable to one-sided tests.

Given the variety of multiple-comparisons procedures available, the simplicity and versatility of the Bonferroni procedure—which works for any *p*-values regardless of how they were obtained—make the Bonferroni procedure useful to teach as a default method of Type I error control (Harris, 2001). However, it is important to note that the Bonferroni procedure controls not only the FWER but also the PFER. Failing to understand this may lead to the confusion discussed in the following section.

Confusion About the Utility of the Bonferroni Procedure

The Bonferroni procedure is often described as "overly conservative" (as noted by Gordon et al., 2007), or as being "improved" through modifications such as Holm's and Hochberg's (see Dickhaus, 2014; Posch & Futschik, 2008; Simes, 1986). This framing is legitimate if the goal is to control the FWER. However, if the goal is to control the PFER, then the Bonferroni procedure is not overly conservative, and hence should not be "improved" by modifications that make it more liberal. Thus, the Bonferroni procedure is perhaps better depicted not as a "blunt tool" (Miles & Banyard, 2007, p. 263) for controlling the FWER—but rather as a precise and efficient tool for controlling the PFER.

Psychology researchers that have touted the superior power of stepwise methods over the Bonferroni procedure (e.g., Blakesley et al., 2009; Eichstaedt et al., 2013; Seaman et al., 1991) have rarely mentioned that such methods—though useful—do not control the PFER and therefore are not adequate substitutes for the Bonferroni procedure when control of the PFER is desired. For example, Eichstaedt and colleagues (2013, p. 693) explicitly stated, "The Holm's sequential procedure corrects for Type I error as effectively as the traditional Bonferroni method"—which is only true if the PFER is not considered (see Barnette & McLean, 2005). Crawley's (2013) canonical guide to statistical analysis in R called the Bonferroni procedure "ridiculously" conservative (p. 534) and stated, "There seems to be no reason to use the unmodified Bonferroni correction because it is dominated by Holm's method, which is valid under arbitrary assumptions" (pp. 534–535). In fact, the inflated PFERs associated with stepwise procedures are so widely unknown among researchers that Klockars and Hancock (1994) were moved to call inflated PFERs "the hidden costs" of stepwise procedures.

In summary, lack of acknowledgment for the PFER may be causing unnecessary

controversy and confusion: Some present the Bonferroni procedure as an appropriate method for controlling the FWER; others present the Bonferroni procedure as underpowered and obsolete; and neither of these opposing views takes into account the procedure's usefulness for controlling the PFER. However, if the Bonferroni procedure were presented as a method for controlling the PFER, then there would be no dissonance between: (1) recommending the Bonferroni procedure for controlling the PFER, and (2) recommending more powerful methods for controlling the FWER.

The PFER May Be More Relevant Now Than in the Past

There was a time when choosing between the FWER and the PFER appeared to be relatively inconsequential. Miller (1966, p. 10) called the choice "essentially a matter of taste," and acknowledged that he preferred the FWER "for feelings he [could not] entirely analyze." Similarly, Tukey (1953, p. 5) wrote that either error rate could be used in practice and that the FWER merely had "theoretical advantages." Indeed, the Bonferroni procedure's maximum FWER is known to be only trivially different from its maximum PFER. However, selecting an error rate is no longer simply an inconsequential matter of personal preference, given the development of procedures—such as the Holm, Hochberg, and Hommel methods—that can control the FWER while allowing considerable inflation of the PFER. The following simulations demonstrate this inflation in multivariate designs (for other designs, see Barnette & McLean, 2005; Klockars & Hancock, 1994; Shaffer et al., 2013).

Methods

Monte Carlo simulations were conducted in R (version 3.0.2; R Core Team, 2013) of two-group designs with 50 subjects per group. Three numbers of outcome variables were used:

$m = 2$, $m = 5$, and $m = 10$. All observations were randomly sampled from a multivariate normal distribution. Equal population correlations (ρ) between outcome variables were set at 200 values between 0 and 1. All effect sizes (i.e., population mean differences) were set at zero so that any statistically significant sample mean difference between groups would be a Type I error. There were 100,000 simulations for each combination of m and ρ. These simulations generated pseudorandom sample mean differences and sample covariance matrices.

Two-sided univariate tests of the sample mean differences were conducted at $\alpha = .05$ using each of the following four procedures: Bonferroni, Holm, Hochberg, and Hommel. For each of these procedures at each combination of m and ρ, the FWER was computed by dividing the number of simulations in which significance occurred by 100,000, and the PFER was computed by dividing the number of significant tests by 100,000.

Results

Figures 2A, 2B, and 2C show the results for $m = 5$, $m = 5$, and $m = 10$, respectively. At each value of m, each of the four procedures had a maximum FWER of .050, but the PFER could differ notably from the FWER when outcome variables were correlated. For example, Figure 2B shows that for five outcome variables, even a "moderate" correlation of .6 inflated the Hommel procedure's PFER to approximately 0.067. In other words, although the chance of making a Type I error in a given family remained less than one in 20, the rate of Type I errors per family was approximately one in 15. The stepwise procedures can allow even greater PFER inflation at higher values of m and ρ, but the Bonferroni procedure's maximum PFER is always equal to α.

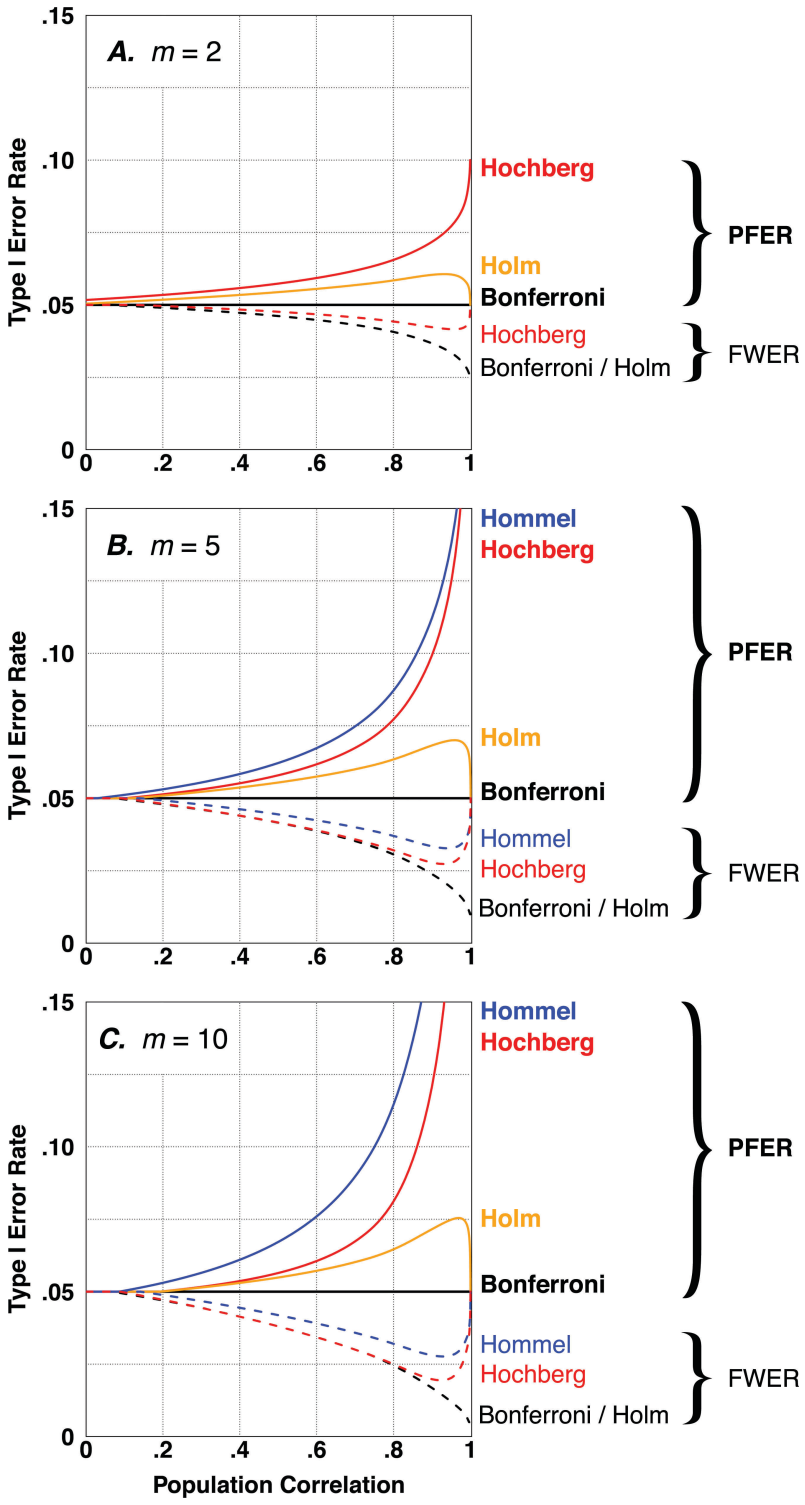

Figure 2. Per-family and familywise Type I error rates for the Bonferroni, Holm, Hochberg, and Hommel procedures in a two-group design with *m* outcome variables (50 subjects per group, α = .05, all null hypotheses true).

Note that in Figures 2B and 2C, the maximum PFERs of the Hochberg and Hommel procedures are well beyond the upper limits of the graphs. At any value of m, the maximum PFER for both procedures approaches $\alpha \times m$ as ρ goes to 1. However, extending the range of the vertical axes to accommodate the extremely inflated PFERs at impractically high correlations would have sacrificed detail in the busier portions of the graphs while adding little useful information.

Discussion

Previous studies (Barnette & McLean, 2005; Klockars & Hancock, 1994; Shaffer et al., 2013) showed that the PFER can be substantially inflated in multigroup designs even when the FWER is controlled. This paper has built on those findings in three principal ways: (1) by demonstrating through simulation that those findings extend to multivariate designs, (2) by graphically illustrating how the population correlation between outcome variables can enhance the disparity between the PFER and the FWER, and (3) by using the applied statistics literature to show that inadequate acknowledgement of the PFER may be causing unnecessary controversy and confusion, particularly with regard to the utility of the Bonferroni procedure.

Implications for Research Practice

This paper proposes that, depending on the research situation, either the PFER or the FWER may be more relevant than the other. Controlling the PFER (i.e., using the Bonferroni procedure) is appropriate when every mistake "hurts"—as is frequently the case in social and behavioral science experiments and in clinical research. For example, if a psychological therapy is found to significantly improve multiple symptoms, then it would be worse for many of those purported improvements to be Type I errors than for only one to be a Type I error. If statistical

power is of concern, then improving the measures and manipulations or increasing the sample size would be a better solution than using a more liberal error rate that increases tolerance of false findings.

Controlling the FWER may be sufficient when, given one Type I error, additional Type I errors are not very costly. Controlling the FWER may also be sufficient when the probability of multiple Type I errors is so low that inflation of the maximum PFER is negligible, as is the case when the number of tests is not exceedingly large and the correlation between tests is known to be near zero (as evident from Figure 2). When strict control of the PFER is not deemed necessary, a method more powerful than the Bonferroni procedure may be used, such as the Hommel procedure (if no confidence intervals are required), or a context-specific method appropriate for the given situation (see Dmitrienko et al., 2010a for an extensive list). A caveat is that the Hochberg and Hommel procedures can fail to control the FWER for negatively correlated one-sided tests (because they are based on Simes, 1986; see Samuel–Cahn, 1996), whereas the Bonferroni and Holm methods do not have this limitation.

Implications for Pedagogy

If the PFER is to be addressed more in practice, then it must also be addressed more in the classroom and in textbooks. Therefore, perhaps professors and textbook authors should include discussion of the PFER along with discussion of the FWER. Additionally, when a multiple-comparisons procedure is described, the specific error rates that it controls (and does not control) should be accurately identified. It is not sufficient simply to refer to "the Type I error rate."

POWER AND TYPE I ERROR CONTROL FOR UNIVARIATE COMPARISONS IN MULTIVARIATE TWO-GROUP DESIGNS

Simulations were used to evaluate statistical power and Type I error control for several multiple-comparisons procedures in multivariate two-group designs. Stepwise procedures, which are known to control the familywise Type I error rate, tended to be more powerful than other methods but did not control the per-family Type I error rate (PFER). Only two methods controlled the PFER: the classical Bonferroni procedure and a type of MANOVA-protection. Which of these two procedures was more powerful depended on multiple factors that this article describes in detail and illustrates graphically. It is concluded that which multiple-comparisons procedure is preferable depends on the number and correlation of outcome variables, the importance of the PFER, the necessity of simultaneous confidence intervals, the satisfaction of assumptions, and the value added when significance is obtained in more than one variable.

Introduction

Experimental designs often examine univariate differences between two groups with respect to more than one outcome variable. For these situations, as noted by Bird and Hadzi–Pavlovic (2014), one of two multiple-comparisons procedures (MCPs) has generally been recommended for controlling Type I errors: either the Bonferroni procedure or a type of MANOVA-protection (MP; described in the section under that name in this paper). However, despite an abundance of conflicting opinions on which method is preferable, previous studies have not thoroughly compared the statistical power of these two procedures. The issue is further complicated by the development of stepwise Bonferroni-based methods that are regarded as

more powerful than the classical Bonferroni procedure.

The present study evaluates several MCPs and addresses relevant perspectives from the literature. Note that the scope of this paper is limited to two-sided univariate comparisons for variables of equal interest, so procedures involving *a priori* weighted hypotheses, contrasts, or combinations among variables are not addressed. The scope is also limited to two-group designs. However, previous studies (Barnette & McLean, 2005; Klockars & Hancock, 1994; Shaffer et al., 2013) have compared power and error rates for MCPs in the context of multiple group comparisons. Moreover, Bird and Hadzi–Pavlovic (2014) found MP to be superfluous for error rate control in designs with more than two groups.

Defining the Type I Error Rate

In multivariate contexts, MCPs can address one or more of three basic Type I error rates (though additional error rates have been proposed; see Dmitrienko et al., 2010a; Hochberg & Tamhane, 1987; Miller, 1966). Which error rate is an appropriate standard depends on the research situation, as summarized in the following sections. In these sections, and throughout this paper, let α be the overall alpha level (i.e., the designated maximum acceptable Type I error rate for a study) and let m be the number of univariate comparisons (which in a two-group design is equal to the number of outcome variables).

Familywise Type I Error Rate (FWER). The FWER is the probability of making a Type I error in at least one of m comparisons. This is the most commonly used Type I error rate in multivariate contexts and tends to be the only error rate mentioned in statistics textbooks (notwithstanding certain texts specifically focused on multiple testing, e.g., Dmitrienko et al., 2010b; Hochberg & Tamhane, 1987; Miller, 1966; Toothaker, 1992). Note that the FWER does not distinguish the situation in which one discovery in a study is erroneous from the situation in

which many discoveries in a study are erroneous. Thus, the FWER is an appropriate standard for situations in which the occurrence of multiple Type I errors is not considered substantially worse than the occurrence of one Type I error. When $m = 1$, the maximum FWER is equal to α. Therefore, the maximum FWER must remain less than or equal to α to be considered uninflated ("controlled") as m increases.

False Discovery Rate (Benjamini & Hochberg, 1995). Designed for situations in which "the FWER is not quite needed" (Benjamini & Hochberg, 1995, p. 290), the false discovery rate is, loosely speaking, the expected proportion of rejected null hypotheses that are true (if the proportion is considered as zero whenever there are no rejections). This can be a useful standard when the number of outcome variables is very large and controlling the FWER would be too costly in power. However, because the false discovery rate is a more lenient standard than the FWER, its applicability may be limited to contexts where the cost of individual Type I errors is relatively low (Dmitrienko et al., 2010a; Finner & Roters, 2001; Goeman & Solari, 2014; Meijer & Goeman, 2016).

Per-Family Type I Error Rate (PFER; Tukey, 1953). The PFER (also called the *expected family error rate*; Miller, 1966, or the *error rate per experiment*; Ryan, 1959) is the expected number of Type I errors out of m comparisons, which is equivalent to the sum of probabilities of Type I error for all m variables. When $m = 1$, the maximum PFER is equal to α. Therefore, the maximum PFER must remain less than or equal to α to be considered uninflated as m increases. Note that although controlling the FWER does not necessarily control the PFER, controlling the PFER always controls the FWER.

Because the PFER is a stricter standard than the FWER, controlling the PFER can require an undesirable sacrifice in power, particularly when m is very large. However, the PFER is

arguably more relevant than the FWER for behavioral science studies in which it is typically worse to make multiple Type I errors than to make only one (Barnette & McLean, 2005; Klockars & Hancock, 1994; Ryan, 1959, 1962). If multiple discoveries are considered to incrementally add value to results, then each "discovery" that turns out to be false must incrementally detract from that aggregate value. It follows that a researcher should not accept credit every time significance occurs without acknowledging a mistake for every error that occurs.

For these reasons, and because the PFER has been examined far less than the FWER, the present study gives considerable attention to the PFER. However, it must be emphasized that none of the three error rates described here should be considered as definitively more appropriate than the others for all situations. In fact, in certain exploratory contexts, formal null hypothesis testing with strict Type I error control may not be necessary at all. The following sections explain the FWER-controlling MCPs examined in the present study, two of which also control the PFER.

The Bonferroni Procedure

If no MCP is used, the maximum Type I error rate for each comparison is equal to α, so the PFER may theoretically be as high as $\alpha \times m$. The Bonferroni procedure prevents PFER inflation by reducing the *testwise alpha level* (the level at which an individual test is conducted) to α / m for each univariate test. Thus, the PFER is capped at $(\alpha / m) \times m = \alpha$.

Stepwise Bonferroni-Based Methods

Šidák Procedure (Šidák, 1967). The Šidák procedure controls the FWER by setting the testwise alpha level at $1 - (1 - \alpha)^{1/m}$. Note that this is only trivially higher than the Bonferroni-adjusted (α / m) level and thus can offer only trivially more power. For example,

when $\alpha = .05$ and $m = 2$, the Bonferroni-adjusted level is .025 and the Šidák-adjusted level is approximately .0253.

Holm Step-Down Procedure (Holm, 1979). Instead of testing all hypotheses at the same testwise alpha level, the Holm procedure performs tests sequentially at different testwise alpha levels: First the smallest *p*-value is compared to α / m, then the next-smallest *p*-value is compared to $\alpha / (m - 1)$, then the next-smallest *p*-value is compared to $\alpha / (m - 2)$, and so on until finally the largest *p*-value is compared to α. After the first test, each test is conditional on significance of the preceding test, meaning that if a *p*-value is nonsignificant at any step, then all larger *p*-values are also considered nonsignificant and testing stops.

Holm's method is more powerful than the Bonferroni procedure for controlling the FWER because Holm's method rejects every null hypothesis that Bonferroni would reject, rejects some null hypotheses that Bonferroni would not, and still controls the FWER. Therefore, many researchers in the health and psychology fields (e.g., Aickin & Gensler, 1996; Eichstaedt et al., 2013; Levin, 1996; Ludbrook, 1998; Seaman et al., 1991; Wright, 1992) have recommended Holm's method as a replacement for Bonferroni adjustment.

Hochberg Step-Up Procedure (Hochberg, 1988). Hochberg's method resembles Holm's but in reverse sequence: First the largest *p*-value is compared to α, then the next-largest *p*-value is compared to $\alpha / 2$, then the next to $\alpha / 3$, and so on until finally the smallest *p*-value is compared to α / m. Conditionality is also reversed relative to Holm's method: With Hochberg's method, each test is conditional on nonsignificance of the previous test, meaning that if a *p*-value is found significant at any step, then all smaller *p*-values are also declared significant and testing stops. Hochberg's procedure is more powerful than Holm's for controlling the FWER because Hochberg's rejects every null hypothesis that Holm's would reject, rejects some null hypotheses

that Holm's would not, and still controls the FWER. Consequently, the Hochberg procedure has been recommended in place of the Bonferroni procedure, e.g., for neuropsychological and pharmaceutical research (Blakesley et al., 2009; Dmitrienko et al., 2010a; Sankoh et al., 1997).

Hommel Procedure (Hommel, 1988). This method is similar to Hochberg's, but is a bit more complex. Hommel's procedure is more powerful than Hochberg's for controlling the FWER because Hommel's rejects every null hypothesis that Hochberg's would reject, rejects some null hypotheses that Hochberg's would not, and still controls the FWER. The algorithm is as follows, where $\{p_1, \ldots, p_m\}$ are the p-values ordered from smallest to largest, and b is the integer vector $\{1, \ldots, m\}$:

> Sequentially, for each value of b from 1 to m, if $p_{(m - j + 1)} < (b - j + 1)\ \alpha / b$ for any $j = \{1, \ldots, b\}$, then $\{p_1, \ldots, p_{(m - b + 1)}\}$ are significant, any other p-values are nonsignificant, and the procedure stops. Else, the procedure continues to the next value of b or, if b has been exhausted, the procedure stops and all p-values are nonsignificant.

Note that the Šidák, Hochberg, and Hommel procedures can allow FWER inflation for negatively dependent one-sided tests, i.e., when two outcome variables are negatively correlated and the hypothesized mean differences are in the same direction for both of them, or when two outcome variables are positively correlated and the hypothesized mean differences are in opposite directions (see Samuel–Cahn, 1996). However, in the conditions examined by the present study (i.e., two-sided tests), the procedures controlled the FWER.

MANOVA-Protection

For designs with two groups, MP is a two-stage process (see Bird & Hadzi–Pavlovic,

2014). In Stage 1, a multivariate analysis of variance (MANOVA) is conducted on all outcome variables simultaneously (equivalent to Hotelling's T^2 in two-group designs). If the MANOVA is significant at level α, then it is followed by Stage 2, in which one univariate test is conducted for each outcome variable. However, if the MANOVA is not significant, then significance of all univariate tests is forfeited. Thus, the MANOVA acts as a *gatekeeper* that can only be "passed" 5% of the time (presuming $\alpha = .05$) when the groups do not truly differ with respect to any of the outcome variables. This paper examines two versions of MP (described as follows), each using a different testwise alpha level adjustment for the univariate tests.

MANOVA-Protection With the α^{**} Adjustment. Bird and Hadzi–Pavlovic (2014) demonstrated that MP controls the FWER for two-sided tests in two-group designs when the following testwise alpha level is used for all univariate tests: $\alpha^{**} = 1 - (1 - \alpha)^{1 / (m - 1)}$, where α^{**} is the testwise alpha level. When $m = 2$, this adjustment reduces algebraically to $\alpha^{**} = \alpha$, meaning that for two outcome variables the univariate tests may be conducted without testwise alpha level adjustment. Because α^{**} is in part based on the Šidák adjustment, it does not necessarily control the FWER for negatively correlated one-sided tests. However, the present study is concerned with two-sided tests.

MANOVA-Protection With the α / (m – 1) Adjustment. Note that α^{**} is equivalent to what the Šidák-adjusted level would be for $m - 1$ comparisons. To convert this adjustment to one that controls the PFER, this paper proposes using the equivalent to what the Bonferroni-adjusted level would be for $m - 1$ comparisons, i.e., $\alpha / (m - 1)$. Like α^{**}, this simpler adjustment is equal to α when $m = 2$. When $m > 2$, it is only marginally lower than α^{**}. This method reliably controls the PFER for $\alpha \leq .05$ (see Appendix A), assuming that the MANOVA test itself is not considered to be part of the family.

Myths About MANOVA-Protection

It is important to distinguish MP from the MANOVA test itself. MANOVA is known to be useful in contexts such as discriminant analysis. MP, on the other hand, has remained controversial for decades (for various perspectives, see Bird, 1975; Bird & Hadzi–Pavlovic, 2014; Bray & Maxwell, 1982; Dar et al., 1994; Enders, 2003; Grice & Iwaski, 2007; Haase & Ellis, 1987; Huberty & Morris, 1989; Huberty & Petoskey, 2000; Hummel & Sligo, 1971; Keselman et al., 1998; Larrabee, 1982; Leary & Altmaier, 1980; Spector, 1981; Share, 1984; Strahan, 1982; and others). Widespread misunderstanding and misapplication of MP may be responsible for much of the disagreement about the procedure. Therefore, the present study's evaluation of MP requires first clarifying the proper implementation of the procedure and correcting some common misconceptions.

Myth 1: MP Never Requires Testwise Alpha Level Adjustment. Many authors (e.g., Enders, 2003; Grice & Iwaski, 2007; Huberty & Morris, 1989; Kellow, 2000; Share, 1984) have identified the common misconception that MP always controls the FWER without testwise alpha level adjustment. In fact, Bird and Hadzi–Pavlovic (2014) noted that in practice MP is typically applied without adjustment. The gatekeeper does cap the FWER at α when all univariate null hypotheses are true, but a real effect in any variable eliminates this protection. Thus, unless $m = 2$ (in which case a real effect in one variable would eliminate the possibility of more than one Type I error), MP requires testwise alpha level adjustment to control the maximum FWER.

Myth 2: MP Requires Classical Bonferroni Adjustment. Just as it would be improper to apply insufficient testwise alpha level adjustment, it would also be improper to apply too much. Recall that MP requires a testwise alpha level of α^{**} to control the FWER or

α / (m – 1) to control the PFER. Each of these testwise alpha levels is higher than the Bonferroni-protected (α / m) level. Thus, although many textbooks (e.g., Harris, 2008; Johnson & Wichern, 2002; Mertler & Vannatta, 2010; Meyers et al., 2006; Spicer, 2005; Stevens, 2002) have recommended that classical Bonferroni adjustment be applied following the MANOVA gatekeeper, this strategy unnecessarily drains power (Bird & Hadzi–Pavlovic, 2014).

Myth 3: MANOVA Is Always a Prerequisite for Univariate Tests When There Are Multiple Outcome Variables. Many believe, as some textbooks have implied (e.g., Christensen, 2007), that an omnibus F-test is always required before multiple mean comparisons are performed—even if the omnibus null hypothesis in itself is not particularly interesting to the researcher. However, there is no apparent justification for such a rule (as noted by Barnette & McLean, 2005; Grice & Iwaski, 2007; Howell, 2013, pp. 372–373; Huberty & Petoskey, 2000; Keselman et al., 1998).

Myth 4: MANOVA Is Incompatible With Univariate Tests. Critics of MP (e.g., Dar et al., 1994; Enders, 2003; Grice & Iwaski, 2007; Huberty & Morris, 1989; Keselman et al., 1998) have rightfully cautioned that MANOVA and univariate tests address different questions: MANOVA compares groups in multidimensional space, whereas univariate tests compare groups on individual variables. However, some authors have gone so far as to declare MANOVA fundamentally "incompatible" with univariate tests (e.g., Enders, 2003, p. 40). Although it is true that MANOVA would be an inadequate substitute for univariate tests, there appears to be no empirical reason that MANOVA should never be used as a precursor to univariate tests.

Note that the MANOVA test statistic may be expressed as a function that includes the univariate test statistics. For example, when samples sizes are equal and m = 2,

$$F = \frac{2n-3}{4n-4} \times \frac{t_X{}^2 - 2rt_X t_Y + t_Y{}^2}{1-r^2},$$

where F is the MANOVA test statistic, n is the sample size per group, t_X and t_Y are the t-statistics for two variables X and Y, and r is the sample correlation between X and Y. Hence, although MANOVA and univariate tests are not interchangeable, they are computationally related.

Myth 5: Correlations of Outcome Variables Predict MANOVA's Power. MANOVA's power is a non-monotonic function of correlations and effect sizes (see Cole et al., 1994). Yet some authors have made generalizations about MANOVA's effectiveness based on correlation alone, stating for instance that MANOVA "works acceptably well" when $|\rho| \approx .6$ and is "less attractive" at certain other correlations (Tabachnick & Fidell, 2012, p. 270). As the present study will illustrate, no particular correlation makes MANOVA especially powerful independently of the effect sizes.

It has been said that MANOVA is more powerful when correlation is highly negative and less powerful when correlation is highly positive (Hair et al., 2006; Meyers et al., 2006; Ramsey, 1982; Tabachnick & Fidell, 2012). However, when $m > 2$, this heuristic is moot because it is mathematically impossible for more than two variables to be highly negatively correlated. Moreover, even when $m = 2$, it is meaningless to distinguish between highly negatively correlated and highly positively correlated outcome variables with respect to MANOVA's power. This is evident from the formula for the MANOVA noncentrality parameter (adapted from Morrison, 1967):

$$\delta^2 = \frac{n}{2} \times \frac{\Delta_X{}^2 - 2\rho\Delta_X\Delta_Y + \Delta_Y{}^2}{1 - \rho^2},$$

where Δ_X and Δ_Y are the standardized effect sizes in Variables X and Y respectively, and ρ is the population correlation between X and Y. MANOVA's power is a monotonically increasing function of noncentrality parameter δ^2. Note that the only part of this formula that is sensitive to the positive/negative directionality of ρ, Δ_X, or Δ_Y is the term $2\rho\Delta_X\Delta_Y$. Note also that the value

of this term is unchanged when the values of any two of the three parameters are multiplied by −1. Hence, for any negative value of ρ, equal power can be obtained for a positive value of ρ simply by reverse-coding Variable X (thereby flipping the signs on ρ and Δ_X) or reverse coding Variable Y (thereby flipping the signs on ρ and Δ_Y). Thus, the heuristic that MANOVA is more powerful for highly negative values of ρ only applies when Δ_X and Δ_Y are both positive or both negative—which would be unlikely if X and Y were highly negatively correlated, as discussed in the following paragraph.

It has long been known (Cole et al., 1994; Morrison, 1967) that MANOVA is particularly powerful when correlation is highly negative and effects are in the same direction. However, with rare exception (Aberson, 2010), authors have neglected to mention that such a combination of parameters (i.e., $\rho\Delta_X\Delta_Y << 0$) is unlikely in practice. Cole et al. (1994) suggested that a training manipulation might increase both speed and accuracy on a task, even though speed and accuracy might be negatively correlated in the population. Other authors (Tabachnick & Fidell, 2012, p. 270; Woodward et al., 1990, p. 392) have proposed similar hypothetical scenarios. However, the example seems improbable; if training increased speed and accuracy, that would suggest that speed and accuracy reflected skill level, meaning that speed and accuracy would likely be positively correlated in the population. That is not to say that situations in which $\rho\Delta_X\Delta_Y << 0$ do not exist, but the absence of cited real-world examples from large-sample studies suggests that such situations are rare.

The Present Study

The present study used Monte Carlo simulations to evaluate the Bonferroni, Šidák, Holm, Hochberg, Hommel, and MP procedures in two-group designs with various *parameter combinations* (distinct combinations of values for the population correlations and standardized

effect sizes of the outcome variables). The MCPs were evaluated in terms of FWER, PFER, *any-variable power* (the probability of finding significance in at least one outcome variable; Ramsey, 1982), and *expected number of significant variables* (NSV, which is equivalent to the sum of powers for all *m* outcome variables). Experiment 1 examined designs with two outcome variables; experiment 2 examined designs with three outcome variables; and experiment 3 examined designs with up to 20 outcome variables. In each of the three experiments, a large number of simulations was used, so that estimated standard error was ≤0.0003 for each error-rate estimation and ≤0.0005 for each power-difference estimation (see Albert & Rizzo, 2012, p. 309).

Experiment 1 – Methods

SAS 9.3 software was used to generate pseudorandom datasets (*simulations*) with bivariate normal outcome variables (X and Y) and two groups ($n = 50$ per group, which is plausible for a small behavioral study in which statistical power could be a concern). Population covariance matrices were equal, which is a standard assumption for MANOVA. Each simulation was performed using assigned values for the following three parameters: Δ_X (standardized effect size in Variable X), Δ_Y (standardized effect size in Variable Y), and ρ (population correlation between X and Y). There were 1,000,000 simulations for each parameter combination.

Parameter Combinations

Without loss of generality, all population standard deviations were fixed at 1. Thus, Δ_X and Δ_Y represented standardized differences in population means. Δ_X was assigned values ranging from −1.2 to 1.2 in increments of 0.01, and Δ_Y was assigned values ranging from 0 to 1 in increments of 0.2. It was not necessary to consider larger effect sizes because results indicated

that (at $n = 50$) power for X neared 100% when $|\Delta_X| > 1$, regardless of which MCP was used. The following values were assigned for ρ: 0, .2, .4, .6, .8, and .95. It was not necessary to use negative values of ρ, because switching the direction ρ would have the same effect on power as switching the direction of Δ_X. In order to thoroughly investigate Type I error rates, additional simulations were conducted for values of ρ between 0 and 1 (in increments of .0001) where at least one effect size was zero. It was not necessary to vary *n* because quadrupling the sample size has approximately the same impact on power as doubling the effect sizes (barring very small *n*).

Computations

Five MCPs were examined: Bonferroni, Šidák, Holm, Hochberg, and MP. It was not necessary to include Hommel's method because it is mathematically equivalent to Hochberg's when $m = 2$. It was not necessary to distinguish between the α^{**} and $\alpha / (m - 1)$ adjustments for MP because both are equal to α when $m = 2$. The univariate tests were two-sided two-sample *t*-tests using the pooled standard deviation. α was set to .05.

For each MCP in each parameter combination, two error rates were computed (FWER and PFER) and two power statistics were computed (any-variable power and NSV). Recall that the number of simulations for each parameter combination was 1,000,000. Thus, FWER = (number simulations in which one or more variables with zero effect size was significant) / 1,000,000; PFER = (total number of significant variables with zero effect size from all simulations) / 1,000,000; any-variable power = (number of simulations in which one or more variables was significant) / 1,000,000; and NSV = (total number of significant variables from all simulations) / 1,000,000. To avoid potentially confusing discontinuities in the graphs, significance was counted toward power even when it also was defined as Type I error (or in the very rare cases when a sample mean difference was in the opposite direction from the true effect

size). Hence, effect sizes of zero in the graphs should be considered as being arbitrarily close to zero rather than true zero.

Experiment 1 – Results

Type I Error Rates

All MCPs controlled the FWER at .050. However, only Bonferroni and MP strictly controlled the PFER; the maximum PFER was 0.050 for Bonferroni and MP, 0.051 for Šidák, 0.061 for Holm, and 0.100 for Hochberg. That said, PFER inflation for Holm and Hochberg was marginal when ρ was small, because multiple Type I errors occurring at once is relatively rare in that situation. For instance, the PFERs for Holm and Hochberg were 0.052 and 0.053, respectively, when $\rho \leq .2$. Figure 3 shows each MCP's PFER as a function of ρ when both effect sizes are zero. Although MP appears slightly conservative in Figure 3, that is because MP's maximum PFER does not occur when both effect sizes are zero, but rather when one effect is very large (essentially assuring passage of the gatekeeper) and the other effect size is zero.

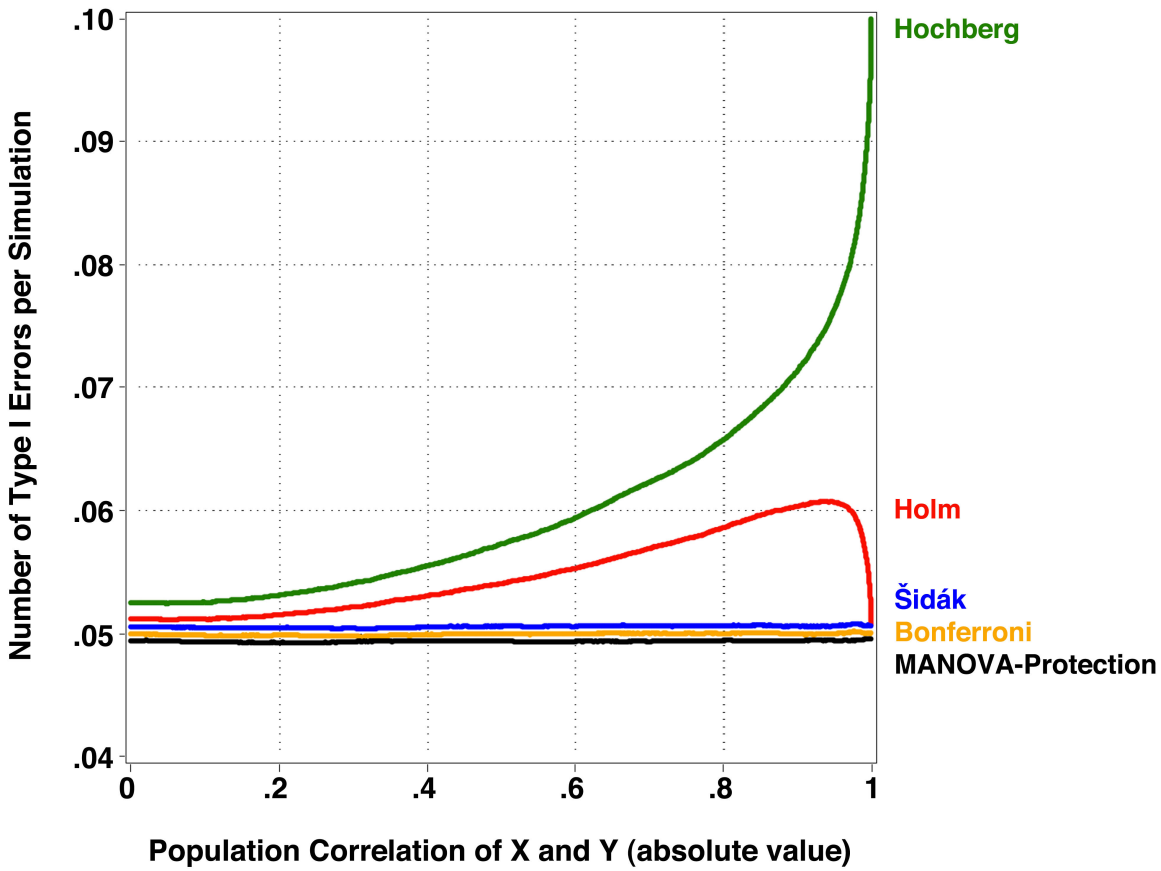

Figure 3. Number of Type I errors per simulation for each multiple-comparisons procedure (2 groups, n = 50 per group, α = .05, 2 outcome variables X and Y, both effect sizes zero).

Bonferroni vs. MANOVA-Protection: Any-Variable Power

Note that the Šidák procedure's power is too trivially different from Bonferroni's to merit a separate discussion. Note also that the Holm procedure's any-variable power is identical to Bonferroni's (provided that significance is always counted toward power) because the two procedures share the same minimum requirement for significance: The lowest p-value must be less than α / m. Thus, all references to Bonferroni in this section can be taken to apply to the Šidák and Holm procedures as well.

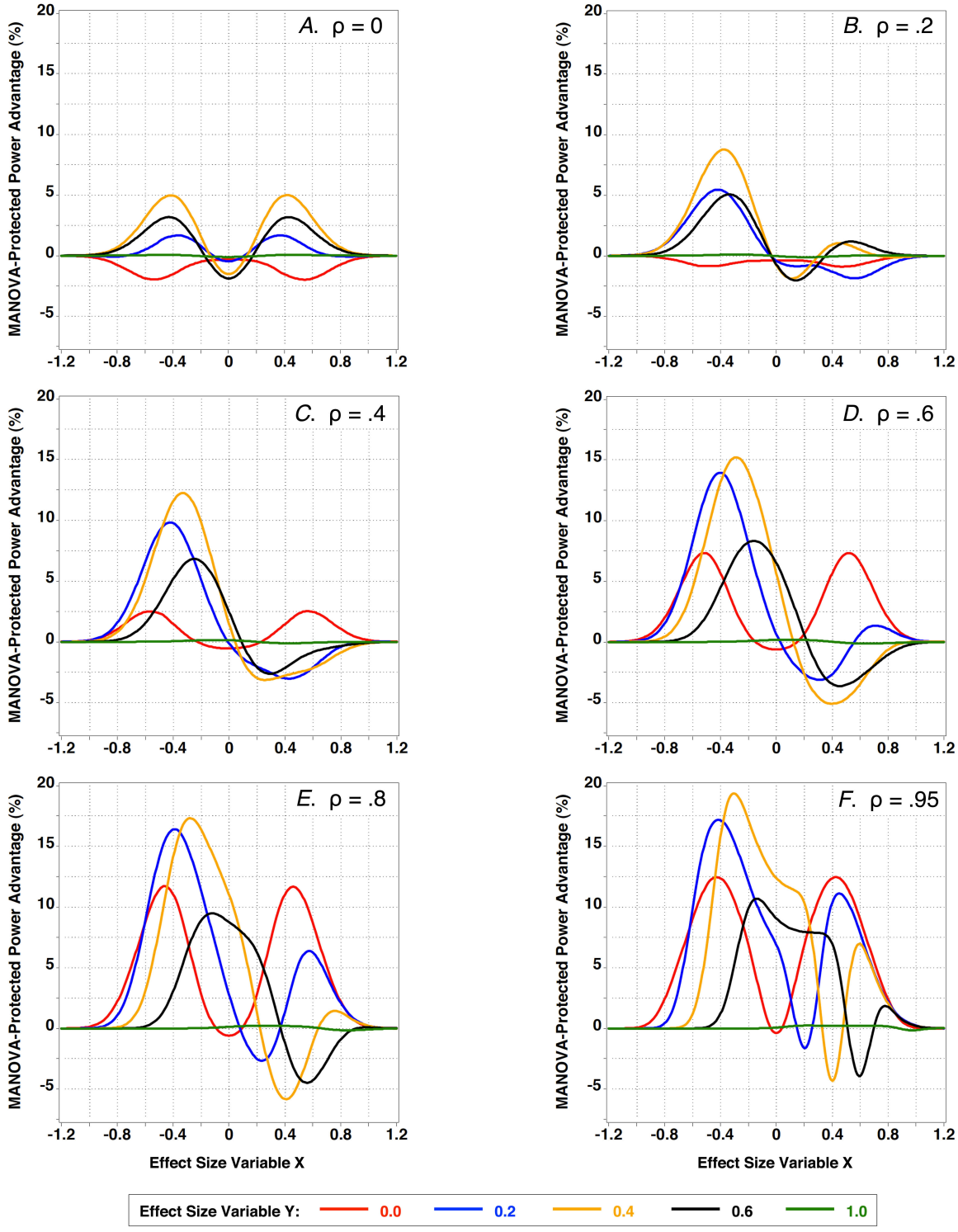

Figure 4. Difference in any-variable power: MANOVA-protection minus Bonferroni (2 groups, $n = 50$ per group, $\alpha = .05$, 2 outcome variables X and Y by their population correlation ρ).

In each of the graphs in Figure 4, the vertical axis indicates the difference in any-variable power. Points above 0% indicate advantages for MP and points below 0% indicate advantages for Bonferroni. Different graphs correspond to different values of ρ. The horizontal axes indicate Δ_X and different colored curves represent different values of Δ_Y.

The competition between Bonferroni and MP is essentially a competition between Bonferroni and the MANOVA gatekeeper: If the gatekeeper is passed, then MP cannot be "beaten" (because MP's testwise alpha level is higher than Bonferroni's), and if the gatekeeper is not passed, then MP cannot "win" (because both variables are forfeited). Thus, regions of advantage for MP reflect MANOVA "strong spots" where the gatekeeper's power is relatively high for the given effect sizes, and regions of advantage for Bonferroni correspond closely (though not exactly) to MANOVA "weak spots" where the gatekeeper's power is relatively low for the given effect sizes. These weak spots occur wherever $\Delta_X = \rho\Delta_Y$ or $\Delta_Y = \rho\Delta_X$ (see Appendix B for details).

Wherever ρ or Δ_Y is zero, curves are horizontally symmetrical around $\Delta_X = 0$. As correlation increases (from Figure 4A to Figure 4F), MP's any-variable power advantage for $\Delta_Y > 0$ increases dramatically on the left sides of the graphs, indicating that in those regions the gatekeeper is passed nearly all of the time. However, readers are advised not to be overly impressed by MP's advantages in these arguably unlikely regions where positively correlated variables are oppositely affected by the same factor. Readers are instead advised to focus primarily on the right side of each graph ($\Delta_X \geq 0$), where parameter combinations are more realistic.

There are regions of advantage for MP on the right sides of the graphs, but they are often

unlikely as well. For example, MP has an advantage where ρ is very small while Δ_X and Δ_Y have substantial and relatively similar effect sizes (see the peaks of the blue, orange and black curves on the right side of Figures 4A and 4B). MP also has an advantage where X and Y are substantially correlated while one effect size is considerable and the other is zero (see the peaks of the red curve in Figures 4C, 4D, 4E, and 4F), or where ρ is high while Δ_X and Δ_Y have very different magnitudes (see the peaks of all curves on the right side of Figures 4E and 4F). These scenarios are possible but unlikely because the more similar two variables are, the more similarly they should be expected to respond to the same factor.

In contrast to MP, Bonferroni appears to have any-variable power advantages in very realistic parameter combinations. For example, Bonferroni has an advantage where ρ is small and effect sizes are not both large (see the troughs of the curves in Figures 4A and 4B). Bonferroni also has an advantage where ρ is moderately positive while X and Y have effects that are at least somewhat similar (see the troughs of the blue, orange, and black curves in Figures 4C and 4D), or where ρ is highly positive while X and Y have effects that are very similar (see the troughs of all curves in Figures 4E and 4F). Note that there are some realistic regions of advantage for MP, but these advantages are relatively small. For example, see the peak of the blue curve on the right side of Figure 4D ($\rho = .6$, $\Delta_X \approx 0.7$, $\Delta_Y = 0.2$).

Bonferroni vs. MANOVA-Protection: Number of Significant Variables

All references to Bonferroni in this section can be taken to apply to Šidák as well. Graphs in Figure 5 are labeled similarly to those in Figure 4, except that the vertical axes indicate the difference in NSV. Points above zero indicate advantages for MP and points below zero indicate advantages for Bonferroni. As in Figure 4, readers are advised to focus on the right side of each graph, where parameter combinations are more plausible.

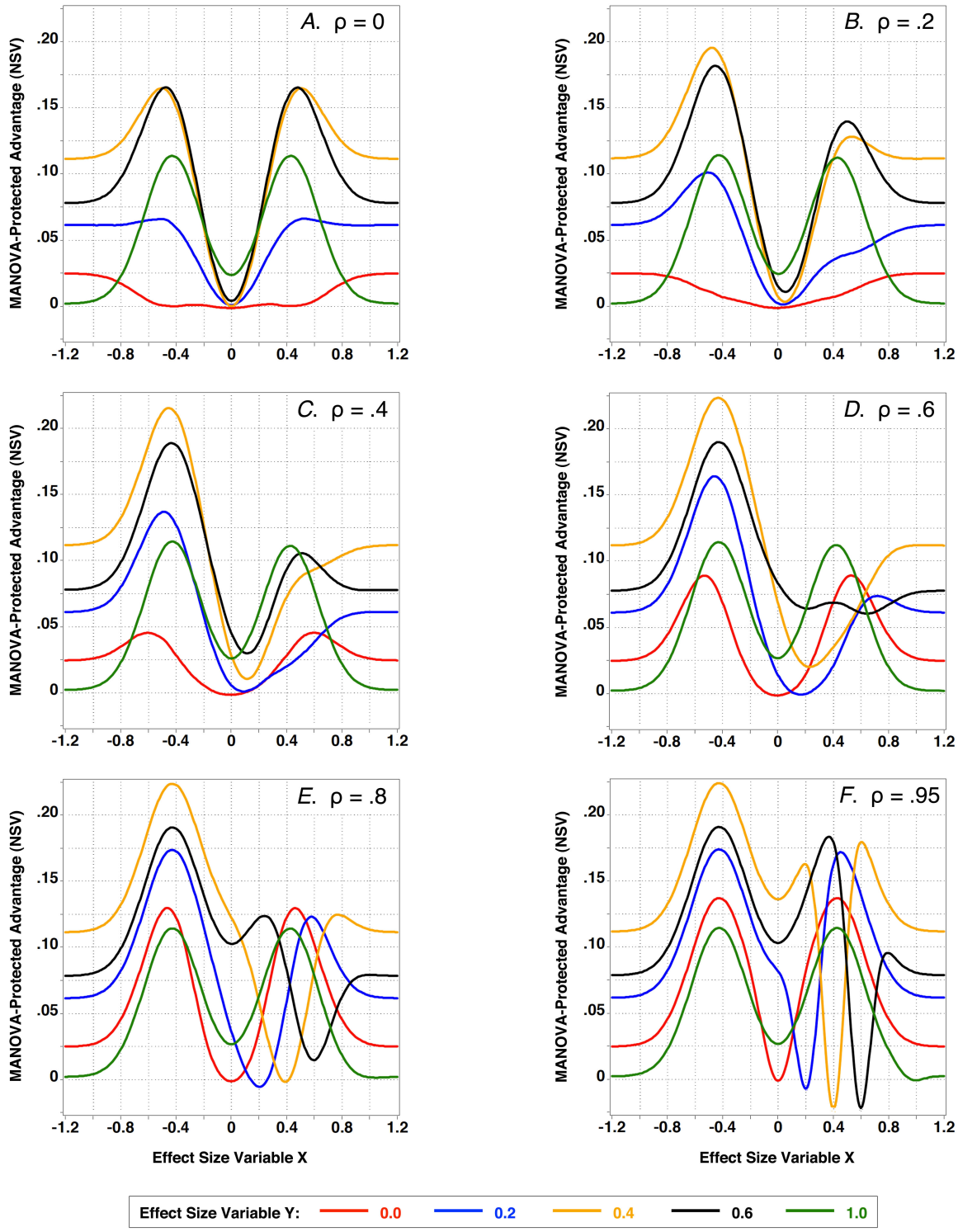

Figure 5. Difference in number of significant variables (NSV): MANOVA-protection minus Bonferroni (2 groups, $n = 50$ per group, $\alpha = .05$, 2 outcome variables X and Y by their population correlation ρ).

Consistent with Figure 4, the right sides of the graphs in Figure 5 show that the troughs of the curves occur in the most realistic situations. However, unlike in Figure 4, the troughs never dip substantially below zero except at the improbably high (.95) correlation. This is not surprising, because it is highly unlikely for the gatekeeper to miss multivariate significance at the .05 level when both univariate tests would be significant at the .025 level. In fact, this occurred in less than 0.02% of simulations, and occurred primarily at $\rho = .95$. Thus, although MP can often lose to Bonferroni at finding significance in one variable, MP is generally superior at finding significance in two variables and consequently produces a greater NSV.

Stepwise Bonferroni-Based MCPs vs. MANOVA-Protection

Any-Variable Power. The any-variable power differences between Šidák and MP, and between Holm and MP, may be taken as equivalent to the any-variable power differences between Bonferroni and MP. The any-variable power differences between Hochberg and MP also showed a similar pattern (e.g., MP's larger advantages occurred in less plausible parameter combinations), but Hochberg performed more favorably against MP than Bonferroni did.

Number of Significant Variables. For Holm vs. MP and Hochberg vs. MP, the patterns of NSV difference were similar to the patterns of any-variable power difference. Hence, the Holm and Hochberg methods often (though not always) produced a higher NSV than MP in realistic parameter combinations. This is a notable distinction from Figure 5, which shows nearly universal NSV advantages for MP over Bonferroni. The reason for this distinction is that with Bonferroni, each variable is tested at a disadvantageous testwise alpha level relative to MP, whereas with Holm and Hochberg, only the variable with the lowest *p*-value is tested at a disadvantageous level relative to MP. In fact, when both univariate *p*-values are less than α,

Hochberg effectively tests both variables at the unadjusted alpha level.

Experiment 2 – Methods

Simulations in Experiment 2 were similar to those in Experiment 1, but with three outcome variables (X, Y, Z) instead of two. Sample size was again fixed at $n = 50$ per group. Values were assigned for standardized effect sizes (Δ_X, Δ_Y, Δ_Z) and for population correlations between outcome variables (ρ_{XY}, ρ_{XZ}, ρ_{YZ}). There were 1,000,000 simulations for each parameter combination.

Δ_X was assigned values ranging from −1.2 to 1.2 in increments of 0.01. Δ_Y and Δ_Z were each assigned values ranging from 0 to 1 in increments of 0.2. ρ_{XY}, ρ_{XZ}, and ρ_{YZ} were each assigned values ranging from 0 to .8 in increments of .2. Certain combinations of correlations were excluded because they were redundant, mathematically impossible, or otherwise unreasonable (e.g., one variable was a linear combination of the other two). In order to thoroughly investigate Type I error rates, additional simulations were conducted for values of ρ_{XY}, ρ_{XZ}, and ρ_{YZ} between −.5 and 1 (in increments of .0001) where at least one effect size was zero. Seven MCPs were examined: Bonferroni, Šidák, Holm, Hochberg, Hommel, MP with the α^{**} adjustment, and MP with the $\alpha / (m - 1)$ adjustment. Computations were analogous to those in Experiment 1.

Experiment 2 – Results

Type I Error Control

All MCPs controlled the FWER at ≤.050. But only the Bonferroni procedure and MP with the $\alpha / (m - 1)$ adjustment strictly controlled the PFER; the maximum PFER was 0.050 for

Bonferroni and MP with the α / (m – 1) adjustment, 0.051 for Šidák, 0.066 for Holm, 0.150 for Hochberg, 0.150 for Hommel, and 0.051 for MP with the α^{**} adjustment. Hence, just as Šidák allowed trivially more Type I errors than Bonferroni while providing trivially more power, α^{**} allowed trivially more Type I errors than α / (m – 1), while providing trivially more power. As in Experiment 1, PFER inflation for the stepwise methods was marginal when ρ was small.

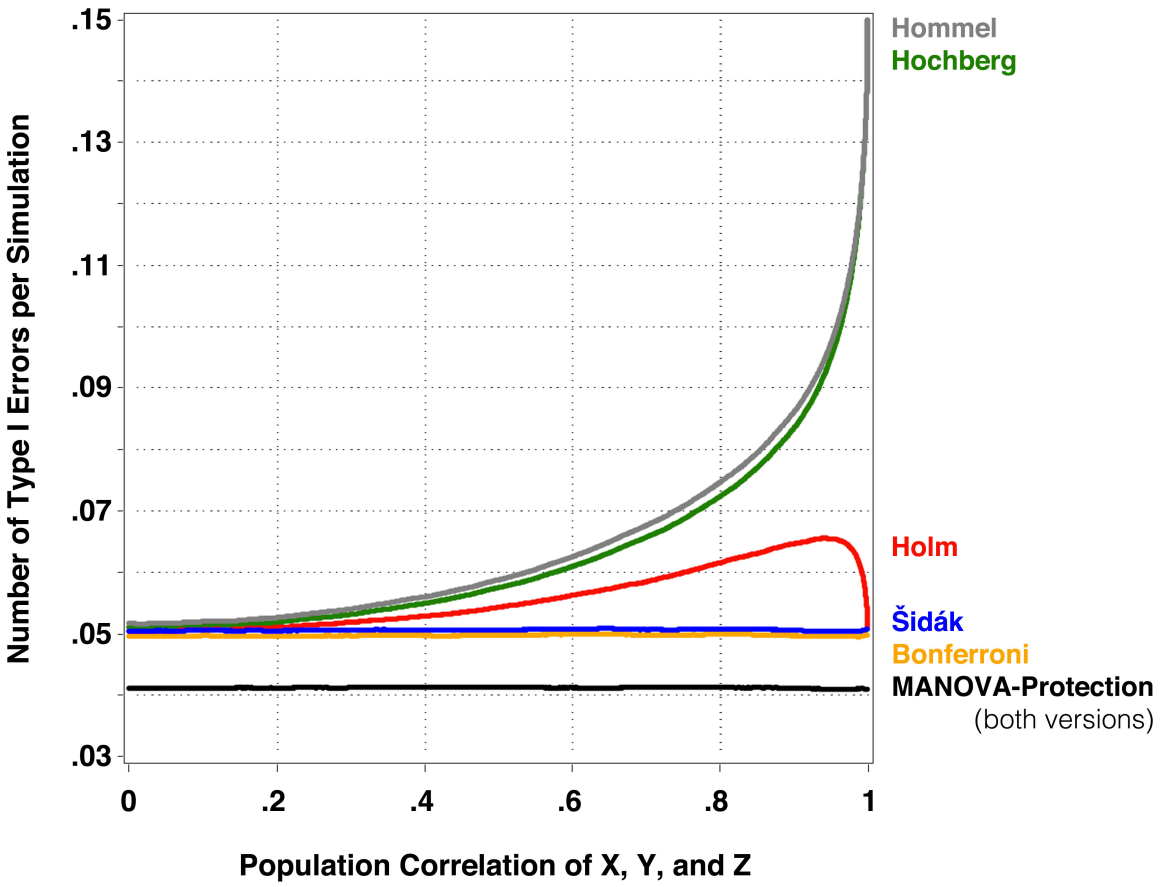

Figure 6. Number of Type I errors per simulation for each multiple-comparisons procedure (2 groups, n = 50 per group, α = .05, 3 outcome variables X, Y, and Z; all effect sizes zero). PFERs for the two versions of MANOVA-Protection, though not strictly identical, are too similar to merit separate lines in the figure.

Figure 6 shows each MCP's PFER as a function of ρ (where $\rho = \rho_{XY} = \rho_{XZ} = \rho_{YZ}$) when all effect sizes are zero. Although MP appears conservative in Figure 4, that is because MP's maximum PFER does not occur when all effect sizes are zero, but rather when there is one very large effect and all other effect sizes are zero.

Bonferroni vs. MANOVA-Protection: Any-Variable Power

Note that all references to Bonferroni in this section can be taken to apply to the Šidák and Holm procedures as well. Note also that the figures discussed in this section were generated based on MP with the $\alpha / (m - 1)$ adjustment but would not be substantially different for MP with the α^{**} adjustment. Compared to Experiment 1, Experiment 2 showed smaller any-variable power advantages for MP and larger any-variable power advantages for Bonferroni. However, results from Experiment 2 were similar to results from Experiment 1 in that MP's any-variable power advantages tended to occur in less realistic parameter combinations and Bonferroni's any-variable power advantages tended to occur in more realistic parameter combinations.

For example, Figure 7A represents a situation where X, Y, and Z are correlated at .6 and $\Delta_Z = 0.4$. In this situation, it is likely that there would be positive effects in X and Y (since X and Y are substantially positively correlated with Z, which has a substantially positive effect size). Hence, the most likely regions of Figure 7A are the blue, orange, and black curves in the right half of the graph—where Bonferroni tends to have a clear any-variable power advantage.

Figure 7B represents a situation where Y and Z are correlated at .6, X is not correlated with Y or Z, and $\Delta_Z = 0.4$. In this situation, it is likely that Δ_X would be small (because X is completely uncorrelated with the affected Variable Z) and it is likely that Δ_Y would be greater than zero (because it is substantially positively correlated with Z). Hence, the most realistic regions of Figure 7B are directly in the middle of the troughs of the blue, orange, and black

curves—again a clear any-variable power advantage for Bonferroni.

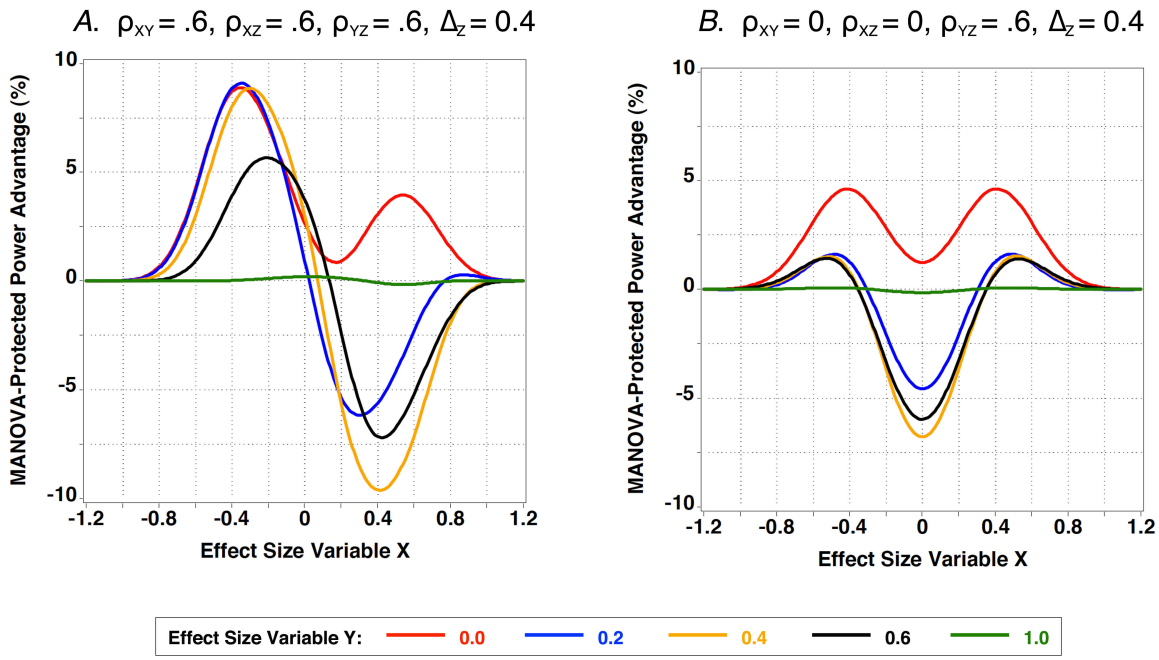

Figure 7. Difference in any-variable power: MANOVA-protection minus Bonferroni (2 groups; n = 50 per group; α = .05; 3 outcome variables X, Y, and Z; Δ_Z indicates standardized effect size for Z; ρs indicate population correlation between the subscripted variables).

Bonferroni vs. MANOVA-Protection: Number of Significant Variables

All references to Bonferroni in this section can be taken to apply to Šidák as well. As in Experiment 1, MP generally produced a higher NSV than did Bonferroni. However, the magnitudes of MP's NSV advantages were notably smaller than in Experiment 1 and Bonferroni had the advantage in certain realistic parameter combinations (albeit by a small margin). Thus, Bonferroni performed better against MP when $m = 3$ than when $m = 2$, in terms of both any-variable power and NSV.

Stepwise Bonferroni-based MCPs vs. MANOVA-Protection

The Holm, Hochberg, and Hommel procedures also performed better against MP when $m = 3$ than when $m = 2$. In fact, the Hochberg and Hommel methods often outperformed MP even in MANOVA strong spots. MP did maintain any-variable power and NSV advantages in certain parameter combinations, but primarily in the improbable situation where one of two highly positively correlated variables had a moderate effect and the other had no effect or had a moderate effect in the opposite direction.

Experiment 3 – Methods

The purpose of Experiment 3 was to ascertain whether Bonferroni's performance against MP would continue to improve as more outcome variables were added. This was accomplished by establishing conservative lower bounds on Bonferroni's maximum advantages and generous upper bounds on MP's maximum advantages for values of m from 2 to 20, as detailed in the following sections. Note that these upper and lower bounds were based on the Bonferroni procedure and on MP with the $\alpha / (m - 1)$ adjustment, but would not be substantially different for Šidák and for MP with the α^{**} adjustment, respectively.

Lower Bounds for Bonferroni's Maximum Advantages Over MANOVA-Protection

For each value of m, systematic searches were conducted for "near-optimal" parameter combinations where Bonferroni's advantages would be maximized. These searches were conducted with an original R (version 3.0.2; R Core Team, 2013) program that used fundamental R functions such as `rnorm`, `rWishart`, `chol`, and `solve`. The program simulated parameter combinations where Bonferroni's advantages were expected to be strong based on the results of

Experiments 1 and 2 (i.e., where substantially and equally correlated variables had moderate and equal effect sizes). Correlations and effect sizes were varied in increments of 0.05. Sample size was fixed at $n = 50$ per group.

Each of Bonferroni's maximum any-variable power advantages occurred where the correlations were between .65 and .75 and the effect sizes were between 0.45 and 0.55. Each of Bonferroni's maximum NSV advantages occurred where the correlations were near one and the effect sizes were between 0.45 and 0.65. Note that the effect sizes at which maximum advantages occurred were specific to $n = 50$, but the values of the maximum advantages would be similar for other sample sizes.

Next, for each of the two power statistics (any-variable power and NSV), 1,000,000 simulations were conducted for each value of m (from 2 to 20). Each simulation compared Bonferroni and MP using the near-optimal parameter combination ascertained for the given power statistic at the given value of m. Resulting power differences served as conservative lower bounds for Bonferroni's maximum advantages over MP.

Upper Bounds for MANOVA-Protection's Maximum Advantages Over Bonferroni

Experiments 1 and 2 showed that MP's maximum advantages (given $n = 50$ and barring $\rho_{XY}\Delta_X\Delta_Y << 0$) occurred where $\rho_{XY} \approx 1$, $\Delta_X > 0.2$, and $\Delta_Y = 0$. These advantageous conditions for MP may be extended to higher values of m by letting all other correlations be zero and all other effect sizes be equal to Δ_X. In these MANOVA strong spots, the gatekeeper's power can be made arbitrarily close to 100% while keeping effect sizes sufficiently moderate to allow MP's testwise alpha level advantage to make a difference at the univariate stage. Under these conditions, systematic searches (using an original simulation program in R, varying Δ_X in increments of 0.05) revealed that MP's maximum any-variable power advantages for $n = 50$

occurred where Δ_X was between 0.25 and 0.40. For each value of *m*, generous upper bounds for MP's any-variable power advantages were then computed by assuming that the gatekeeper was passed 100% of the time and comparing the power of univariate tests at the α / m level to the power of univariate tests at the $\alpha / (m - 1)$ level, using the near-optimal value of Δ_X ascertained for the given value of *m*.

Additional systematic searches (again through an original simulation program in R, varying Δ_X in increments of 0.05) revealed that MP's maximum power advantages per variable for $n = 50$ occurred where the effect size was between 0.45 and 0.60. Given that the NSV is equivalent to the sum of powers for all *m* variables, a generous upper bound for MP's maximum NSV advantage was computed for each value of *m* by simply assuming that the gatekeeper was passed 100% of the time and multiplying *m* by the power advantage per variable (at the near-optimal value of Δ_X ascertained for the given value of *m*).

Experiment 3 – Results

Tables 2 and 3 show that as *m* increased, Bonferroni's maximum advantages increased considerably and MP's maximum advantages decreased considerably—even though the tables use conservative estimates of Bonferroni's advantages and generous estimates of MP's advantages. These results are not surprising. After all, as *m* increases, the difference between α / m and $\alpha / (m - 1)$ becomes progressively smaller, and the ratio between the two adjustments becomes progressively closer to one. Thus, MP's higher testwise alpha level—its only source of power advantage over the Bonferroni procedure—becomes increasingly trivial as *m* increases. For instance, when $\alpha = .05$ and $m = 20$, a univariate *p*-value must fall between .0025 and approximately .0026 to be significant with MP and not with Bonferroni. Moreover, near

MANOVA weak spots, every added variable increases the number of variables at risk of forfeit by the gatekeeper—while costing the gatekeeper in denominator degrees of freedom. Without a strong testwise alpha level advantage at Stage 2 to compensate for this vulnerability at Stage 1, MP becomes far less competitive, in terms of both any-variable power and NSV. Note that the Holm, Hochberg, and Hommel procedures would also improve in performance against MP with every added outcome variable.

Table 2

Maximum Any-Variable Power Advantages for Bonferroni and MANOVA-Protection Over Each Other (m = number of outcome variables)

m	Max. Any-Variable Power Advantage Bonferroni/Holm (Lower Bound)	Max. Any-Variable Power Advantage MANOVA-Protection (Upper Bound)
2	5.7%	12.3%
3	10.6%	7.8%
4	14.5%	5.8%
5	17.9%	4.6%
6	20.6%	3.9%
7	23.6%	3.4%
8	26.0%	3.0%
9	28.2%	2.7%
10	30.2%	2.4%
20	44.9%	1.2%

Table 3

Maximum Number-of-Significant-Variables (NSV) Advantages for Bonferroni and MANOVA-Protection Over Each Other (m = number of outcome variables)

m	Maximum NSV Advantage Bonferroni (Lower Bound)	Maximum NSV Advantage MANOVA-Protection (Upper Bound)
2	.02	.23
3	.05	.19
4	.08	.17
5	.10	.16
6	.12	.15
7	.13	.15
8	.15	.15
9	.16	.14
10	.18	.14
20	.30	.13

Discussion

For controlling the FWER, stepwise Bonferroni-based MCPs (e.g., Holm, Hochberg, and Hommel) are known to be more powerful than the classical Bonferroni procedure, especially as the number of outcome variables increases (though the Holm and Bonferroni procedures have equal any-variable power when all null hypotheses are false). Results of the present study show that stepwise MCPs are often more powerful than MP as well (though not always), especially as the number of outcome variables increases and especially in more realistic parameter combinations. Thus, when FWER is the primary Type I error rate of interest, a procedure other

than Bonferroni or MP may be preferable.

It should be noted, however, that the power advantages of the stepwise procedures come at a cost: inflated maximum PFER. Indeed, the more powerful a given stepwise procedure is, the more Type I errors it potentially allows. Mention of this "hidden" error rate inflation (Klockars & Hancock, 1994) should accompany any future references to the superior power of these techniques. And researchers who use these methods should consider why they are not treating multiple simultaneous false discoveries as substantially worse than a single false discovery. In fact, given the variety of error rates and MCPs to choose from, perhaps any experimenters interested in multiple hypotheses should justify *a priori* why their chosen strategy for controlling Type I errors is appropriate for the context. For example, the PFER is suitable for a study in which the veracity of each claim about a relationship between variables is important (as is likely for many studies intended for publication). The FWER is suitable when the multiple Type I errors are not much worse than a single Type I error, or when the difference between FWER and PFER is negligible (as is the case when m is fairly small and ρ is known to be near zero). The false discovery rate could be suitable when screening through large numbers of variables to determine which should be pursued in a follow-up study.

Results of the present study and of studies addressing the PFER in other experimental designs (Barnette & McLean, 2005; Klockars & Hancock, 1994; Shaffer et al., 2013) demonstrate the need to more adequately acknowledge the PFER. Of the procedures examined, only two strictly controlled the PFER: the Bonferroni procedure and MP with the $\alpha / (m - 1)$ adjustment, though PFER inflation was small when ρ was near zero. Although the PFER inflation allowed by the Šidák and α^{**} methods may be too trivial to be considered problematic, the potential power advantage of these adjustments over the Bonferroni and $\alpha / (m - 1)$ methods

(respectively) is similarly trivial.

For small numbers of outcome variables with plausible parameter combinations, Bonferroni was often (though not always) more likely than MP to produce significance in at least one variable, whereas MP tended to produce a greater NSV than Bonferroni. As more outcome variables were added, Bonferroni's advantages over MP substantially increased while MP's advantages over Bonferroni became increasingly negligible. Thus, when control of the PFER is desired for univariate comparisons in two-group designs, Bonferroni appears generally appropriate. MP (with the α / (m – 1) adjustment) may be preferable when the number of outcome variables is very small and the value of results is substantially incremented with each additional effect detected, but the Bonferroni procedure has certain unquantifiable advantages over MP even in that situation. For example, unlike the Bonferroni and Šidák methods, MP has no associated procedure for computing confidence intervals (Bird & Hadzi–Pavlovic, 2014), a limitation that generally applies to the Hochberg and Hommel methods as well and, except in certain cases, also applies to the Holm procedure (Guilbaud, 2012). This is relevant because a parameter estimate without a confidence interval is typically of limited use—how can one meaningfully interpret an estimate without considering its precision?

Assumption violations may also be of concern. The Bonferroni and Holm procedures require no assumptions beyond those involved in computing the *p*-values. However, MANOVA has substantial additional assumptions (specifically, multivariate normality and homogeneity of covariance) that may make robustness an issue for MP. And although simulations (e.g., Hakstianet al., 1979) have suggested that two-group MANOVA is fairly robust to heterogeneity of covariance in certain situations, robustness has not be definitively established. Recall also that the Šidák, Hochberg, Hommel, and α** methods can fail to control the FWER when the tests are

negatively correlated, though that is typically not plausible for two-sided tests (Samuel–Cahn, 1996).

In short, which MCP is preferable depends on contextual factors: the number and correlation of outcome variables, the importance of PFER versus FWER, the necessity of simultaneous confidence intervals, the satisfction of assumptions, and the relative importance of any-variable power vs. NSV. Therefore, researchers are advised to avoid dogmatic, acontextual stances for or against one method or another. As Grayson (2004, p. 102) analogized, "No carpenter would ever make a statement such as 'hammers are good tools'; they are for some applications, but not for others."

SOME CLARIFICATIONS REGARDING POWER AND TYPE I ERROR CONTROL FOR PAIRWISE COMPARISONS OF THREE GROUPS

A previous study used Monte Carlo simulations to compare the power and familywise Type I error rates of ten multiple-testing procedures in the context of pairwise mean-comparisons in balanced three-group designs. The authors concluded that, of those ten procedures, the Benjamini–Hochberg procedure (BH) was the "best." However, they did not compare BH to classical, commonly used multiple-testing procedures that were developed specifically for pairwise comparisons, such as Fisher's protected least significant difference procedure and Tukey's honest significant difference procedure. Simulations in the present study show that in the three-group case, Fisher's method is more powerful than both Tukey's method and BH, in terms of both *per-pair power* (mean probability of significance across the tests of false null hypotheses) and *any-pair power* (probability of significance in at least one test of a false null hypothesis). Compared to BH, Tukey's method is shown to have lower per-pair power, but slightly greater any-pair power. When population variance was equal, the maximum familywise Type I error rate of all three procedures (BH, Fisher, and Tukey) was equal to the designated alpha level.

Introduction

One of the most common types of statistical analyses for experimental designs is pairwise comparison of group means. Often there are more than two groups, and hence more than one

comparison. In that case, a multiple-testing procedure is typically required to prevent Type I error inflation. Presumably, in practice the most commonly encountered number of groups, besides two, is three. So it is worth asking which multiple-testing methods are preferable in the three-group case. The following three sections provide an inexhaustive list of well known multiple-testing methods that can be used in the three-group case to control the *familywise Type I error rate* (FWER; the probability of at least one Type I error). These methods can typically be executed with simple commands in standard statistical software.

General-Purpose FWER-Control Procedures

Most FWER-control procedures can be applied not only to pairwise mean-comparisons in particular, but also to multiple-testing situations more generally. The best-known such method is the Bonferroni procedure. It controls not only the FWER, but also the *per-family Type I error rate* (PFER; the expected number of Type I errors), which is a stricter standard than FWER. Consequently, the Bonferroni procedure is often regarded as overly conservative, given that FWER has become a preferred standard over PFER in practice (but see Barnette & McLean, 2005; Frane, 2015a, 2015c; Klockars & Hancock, 1994). By sacrificing PFER control, other general-purpose FWER-control procedures (e.g., Holm, 1979; Hochberg, 1988; Hommel, 1988) can provide greater *power* (probability or frequency of significance in tests of false null hypotheses) than the Bonferroni procedure.

The reason FWER control is more “powerful" than PFER control is that, unlike PFER, FWER does not count all Type I errors that occur. Instead, in a given *family* (e.g., in a given study), multiple co-occurring errors count the same as a single error. For instance, if 5 out of every 100 families each contained a single error, then the PFER and FWER would both be .05, but if 5 out of every 100 families each contained 10 errors, then the FWER would still be .05—

even though the PFER would be inflated to .50. Thus, the more an FWER-control procedure can relax protection against co-occurring errors, the more frequently significances can occur, and hence the more powerful the procedure can be while still controlling the FWER.

FWER-Control Procedures Specifically for Pairwise Comparisons

Given that the power of FWER-control procedures is largely based on avoiding "overprotection" against co-occurring errors, knowing how co-occurring errors behave for the specific type of tests at hand allows for the design of more powerful procedures. For instance, for pairwise comparisons of group means, there is an inherent logical relationship between the null hypotheses (e.g., for three group means $\{\mu_1, \mu_2, \mu_3\}$, if $\mu_1 = \mu_2$ and $\mu_2 = \mu_3$, then $\mu_1 = \mu_3$). Additionally, there is correlation among the test statistics in pairwise comparisons, which increases the probability of co-occurrence among errors (e.g., the test of μ_1 versus μ_2 is correlated with the test of μ_1 versus any other mean). By using such information, FWER-control procedures that are designed specifically for pairwise comparisons can provide more power than general-purpose FWER-control procedures.

Tukey's (1953) *honest significant difference procedure* (HSD) is one of the oldest and most widely used methods of FWER control that has been designed for pairwise comparisons (Hancock & Klockars, 1996; Ramsey & Ramsey, 2008). It requires equal group sizes, but there are well known modifications of HSD that accommodate unequal group sizes (e.g., the Tukey–Kramer procedure; Kramer, 1956; Tukey, 1953). HSD also assumes that the population distributions have equal variance, but there are well known modifications that accommodate unequal variance (e.g., Games & Howell, 1976).

An even older procedure for pairwise comparisons is Fisher's (1935) *protected least significant difference procedure* (PLSD), which has been found to be more powerful than HSD

(Ramsey, 1978; Seaman et al., 1991). PLSD works as follows: If the omnibus test (an ANOVA in classical PLSD, though other tests can be used) is statistically significant at the unadjusted alpha level, then the pairwise comparisons (Student *t*-tests in classical PLSD, though other tests can be used) are conducted without adjustment; otherwise, significance of all pairwise comparisons is forfeited. In general, PLSD only controls the FWER "in the weak sense," meaning that it controls the FWER when all null hypotheses are true, but can fail to control the FWER when only some null hypotheses are true (Hochberg & Tamhane, 1987, p. 3). However, procedures that generally only control the FWER in the weak sense also control the FWER "in the strong sense" (meaning regardless of what proportion of the null hypotheses are true) in a special case: pairwise comparison of three groups on a single outcome variable (Hayter, 1986; Shaffer, 1986). To understand why, consider the three possible mean configurations in the three-group case: (1) all means are equal ($\mu_1 = \mu_2 = \mu_3$), in which case weak control of the FWER is as effective as strong control because all null hypotheses are true; (2) only two means are equal ($\mu_1 = \mu_2 \neq \mu_3$), in which case there is no multiple-testing problem, because there is only one true null hypothesis and thus the opportunity for Type I error is confined to a single comparison; (3) all means are different, in which case Type I error is impossible because there are no true null hypotheses. In other words, in the three-group case, multiple-testing can inflate the FWER only when all null hypotheses are true, so the distinction between weak and strong control of the FWER is moot. Thus, although PLSD is not typically advisable when there are more than three groups, it is valid for controlling the FWER in the three-group case, provided that the assumptions of the omnibus test are met.

Other procedures for pairwise comparisons have been developed, but are beyond the scope of this paper. For example, the Newman–Keuls procedure (Keuls, 1952; Newman, 1939)

controls the FWER under certain conditions, but it has been found to be slightly less powerful than PLSD in the three-group case (Seaman et al., 1991). Another method is Dunnett's (1955) procedure, which is applicable not when conducting "all possible" pairwise comparisons, but rather when comparing multiple treatment groups to a single control group (and not comparing the treatments to each other).

The Benjamini–Hochberg Procedure

The Benjamini–Hochberg procedure (BH; Benjamini & Hochberg, 1995) is a general-purpose multiple-testing procedure that was designed to control the *false discovery rate* (the expected "proportion," loosely speaking, of significances that are Type I errors). Because false discovery rate is a more lenient standard than FWER, BH does not reliably control the FWER in general. However, like PLSD, BH controls the FWER when all null hypotheses are true, and therefore controls the FWER in the three-group case. BH has been shown to be valid for independent tests and for many typically-encountered types of positively dependent tests (Benjamini & Yekutieli, 2001; note that negative dependence is not plausible in typical two-sided testing scenarios).

The Félix and Menezes study

Félix and Menezes (FM; 2018) used Monte Carlo simulations to rank the performance of ten multiple-testing procedures in the context of pairwise mean-comparisons in balanced three-group designs. The pairwise comparisons were Student *t*-tests using the pooled standard deviation from all three groups. And the three population distributions were either all normal, all logistic, or all Gumbel. Because BH tended to rank highly in terms of both FWER-control and power, FM concluded that "the BH correction was the best overall, that is, it was good in both criteria" (p. 88). However, that conclusion requires some important caveats.

First of all, FM neglected to mention that BH fails to reliably control the FWER (except in the weak sense) when there are more than three groups. Recall that the FWER control BH provides in the three-group case is based on the fact that there can only be multiple true null hypotheses when all means are equal. When there are more than three groups, there can be multiple true null hypotheses even when not all means are equal (e.g., when $\mu_1 = \mu_2 = \mu_3 \neq \mu_4$).

Another caveat is that FM did not compare BH to classical, commonly used FWER-control procedures that were devised explicitly for pairwise comparisons—most notably, HSD and PLSD. Instead, they compared BH to general-purpose multiple-testing procedures (e.g., Bonferroni), most of which were well known to be less powerful than BH (notwithstanding the Li procedure, which can be liberal for dependent tests; Li, 2008). Thus, for pairwise comparisons of three groups, although BH may indeed often be the best choice among the procedures that FM examined, that does not imply that BH is better than the standard procedures that are available for pairwise comparisons.

Interestingly, despite endorsing BH for pairwise comparisons of group means, FM claimed that BH is only valid for independent tests (p. 79). That claim, if it were true, would invalidate BH for pairwise comparisons of group means. However, although there are indeed some types of dependent tests for which BH is invalid, FM's own results suggest that two-sided pairwise comparison of three group means is not such a type (see also Benjamini & Yekutieli, 2001).

One conclusion in the FM study appears to have resulted from a methodological inconsistency in the simulations. Specifically, FM reported that for the logistic distribution, "the empirical power is a lot smaller, since this distribution has heavy tails" (p. 84). However, as shown in their Table 1, the standard deviations they used for the logistic distribution were larger than the standard deviations they used for the normal distribution by a factor of $\pi / \sqrt{3} \approx 1.8$, and

were larger than the standard deviations they used for the Gumbel distribution by a factor of $\pi / \sqrt{2} \approx 1.4$. Thus, the dramatically reduced power that FM observed for the logistic distribution essentially reflects the standard deviations that were used—not some inherent characteristic of the logistic distribution's shape, such as its slightly "heavy tails." Note that although FM did match the nominal values of the *scale parameters* across the different distribution types, that does not cause the actual spread of the distributions to be matched in any meaningful way, because the scale parameters were defined differently (i.e., as different linear functions of the standard deviation) for the different distribution types.

There are also some issues with how FM ranked the FWER control of the procedures. First of all, FWER was only examined when all population means were equal, i.e., when all null hypotheses were true. Although some procedures (such as Bonferroni) produce their maximum FWER when all null hypotheses are true, other procedures (such as BH) produce their maximum FWER not when all null hypotheses are true, but rather when only some null hypotheses are true and power is maximal (Finner & Roters, 2001). Note also that FM ranked each procedure's FWER control not by how low the FWER was, but rather by how close the FWER was to .05. That approach produces counterintuitive rankings. For instance, using that system, a controlled FWER of .040 would be considered "worse" than an inflated FWER of .059. That explains why, in certain conditions, FM ranked the Li procedure as number 1 in FWER control—even though it was the one procedure in the study that did not actually control the FWER (see their Figure 1). Moreover, for procedures that did control the FWER, higher FWERs were ranked as "better" than lower FWERs.

Perhaps the reason that FM reported rankings for the estimated FWERs rather than reporting the estimates directly is that the number of simulations—only 10,000 for each

combination of parameters—did not provide sufficient precision. Indeed, when using simulations to estimate an incidence rate (such as FWER or power), the standard error for that estimation is inversely proportional to the square root of the number of simulations, as reflected in the following well-known formula: $\widehat{\mathrm{SE}} = \sqrt{v(1-v)/k}$, where $\widehat{\mathrm{SE}}$ is the estimated standard error, v is the observed incidence-rate, and k is the number of simulations (adapted from Albert & Rizzo, 2012, p. 309). The corresponding 95% confidence interval may be computed as $v \pm 1.960\widehat{\mathrm{SE}}$. For instance, when 10,000 simulations collectively produce an estimated FWER of .050, the $\widehat{\mathrm{SE}}$ for that estimation is roughly .002, and the width of the corresponding 95% confidence interval is .009—which is presumably too wide for adequately estimating values beyond two decimal places.

The same principle of precision applies to the power estimates, for which $\widehat{\mathrm{SE}}$ can be as high as .005 (maximized thusly when $v = .5$) and for which the corresponding 95% confidence interval can be as wide as .020. In fact, it is evident that the power estimates in the FM study were noisy, because in some cases, different parameter combinations that should have been computationally equivalent nonetheless produced different power rankings for the same procedure. Examples of this inconsistency can be seen in their Figure 6, which shows that for group size 3 and standard deviation 1, the ranking of the Finner procedure typically changed when the positive/negative sign of the *location parameter* was flipped, i.e., when the directions of the nonzero mean-differences were reversed and the absolute magnitudes of the mean differences were unchanged. Given that the *t*-tests were two-sided, reversing the directions of the mean differences across the board should not have affected power at all. Note also that power was estimated only when two means were equal, so the all-means-different case was not considered.

Despite the FM study's methodological limitations, its primary empirical finding is sound: BH is more powerful than some other procedures that control the FWER for pairwise comparisons of the three groups. Thus, the FM study's main limitation is simply the lack of consideration given to standard procedures that were designed specifically for pairwise comparisons.

The Present Study

The present study followed up on the FM results by conducting Monte Carlo simulations to evaluate the performance of HSD, PLSD, and BH, for pairwise mean-comparisons in balanced three-group designs. This investigation essentially placed BH—the “winning" procedure from the FM study—in competition against classical procedures that were designed for pairwise comparisons. Various group-sizes, population distribution types, and standardized population-mean configurations were used. To avoid the aforementioned methodological limitations of the FM study, the following four steps were taken: (1) standard deviation was fixed, so that power comparisons among distribution types would be meaningful; (2) FWER was estimated not only when all means were equal, but also when only two means were equal (including scenarios in which power was maximal), in order to ensure that each procedure's maximum FWER was produced; (3) power was estimated not only when two means were equal, but also when all means were different; (4) a much larger number of simulations per parameter combination was performed than in the FM study, in order to increase the precision of the estimations.

Methods

Simulation Parameters

Simulations were conducted to evaluate the FWER and power of BH, HSD, and PLSD, in the context of pairwise comparisons of three groups. In each simulation, a group of independent observations was randomly sampled from each of three populations. Group size, which was common to all three groups, was set to 5, 10, 15, 20, or 1000 observations. Population distribution type, which was also common to all three groups, was set to normal, logistic, or Gumbel (following FM). The array of population means was set to {0, 0, 0}, {0, 0, 1}, or {0, 1, 2} ("all-means-equal," "two-means-equal," or "all-means-different," respectively). Population standard deviation was fixed at 1 in all cases, so the population means may be considered as standardized.

1,000,000 simulations were performed for each parameter combination, i.e., for each of the 45 unique combinations of group size, population distribution type, and population-mean array. This is 100 times the number of simulations per parameter combination that was used by FM, and thus provides essentially 10 times the precision (recall that the standard error and the width of the confidence interval are inversely proportional to the square root of the number of simulations). For instance, given an estimated FWER of .050, the 95% confidence interval for the FWER would span from .0456 to .0544 in the FM study, but would span from .0496 to .0504 in the present study—providing tight lower and upper bounds that both round to .050.

Note that parameter manipulations that would be largely redundant with respect to power were not performed in this study. For example, changing the population variance or changing the size of the nonzero population-means would have a similar effect on power as changing the group size. And as previously noted, reversing the positive/negative direction of the population

mean differences across the board would have no effect whatsoever.

Pairwise Comparisons

For BH and PLSD, the pairwise tests were two-sided Student *t*-tests using the pooled standard deviation from the three groups (following FM). For HSD, the pairwise tests (which are essentially built into the procedure) are analogous to those *t*-tests in that they are two-sided, use the pooled standard deviation, and involve the same statistical assumptions (e.g., normality and equal variance).

The familywise alpha level was set to .05 for all procedures. Thus, significance was determined by computing adjusted *p*-values (using the given procedure) and assessing whether they were lower than .05. Note that although PLSD is not typically described as a *p*-value adjustment, it may nonetheless be implemented as such by adjusting the *p*-value for each *t*-test to $\max\{p_t, p_{omni}\}$, where p_t is the raw *p*-value for the given *t*-test and p_{omni} is the *p*-value for the omnibus test.

FWER Estimations

FWER was estimated for each procedure in each parameter combination. The estimated FWER was simply the proportion of simulations that produced at least one Type I error (i.e., the proportion of simulations in which at least one pairwise comparison of groups with equal population means was significant).

Power Estimations

Two types of power were estimated for each procedure in each parameter combination (except when all means were equal, in which case power would be meaningless). *Per-pair power* (the mean probability of significance among comparisons for which the corresponding population mean difference is nonzero; Ramsey, 1978) was estimated by taking the comparisons

for which the corresponding population means were unequal, and computing the proportion of those comparisons that produced significance. *Any-pair power* (the probability of obtaining significance in at least one pairwise comparison for which the corresponding population mean difference is nonzero; Ramsey, 1978) was estimated as the proportion of simulations that produced significance in at least one pairwise comparison of groups with unequal population means.

Software

All simulations and estimations were performed using a custom R program that was created using R version 3.3.3 (R Core Team, 2017). That program, the code for which is provided in Appendix C, contains a section entitled "Adjustable Parameters" that allows the user to specify the following 10 inputs: (1) number of simulations, (2) group size, (3) population mean for Group 1, (4) population mean for Group 2, (5) population mean for Group 3, (6) population standard deviation for group 1, (7) population standard deviation for group 2, (8) population standard deviation for group 3, (9) familywise alpha level, and (10) population distribution type (either "normal," "logistic," or "Gumbel"). Specifying "Gumbel" as the population distribution type requires the `evd` package (Stephenson, 2002), which includes the `rgumbel` function used to generate randomly sampled observations from a Gumbel distribution. All other operations in the program use endogenous R commands, such as `rnorm` (to generate randomly sampled observations from a normal distribution), `rlogis` (to generate randomly sampled observations from a logistic distribution), `aov` (to fit the ANOVA model used as the first step in both HSD and PLSD), `pairwise.t.test` (to perform the pairwise *t*-tests used for both PLSD and BH), `TukeyHSD` (to compute HSD-adjusted *p*-values from the ANOVA model), and `p.adjust` (to compute BH-adjusted *p*-values from the raw *p*-values obtained in the *t*-tests).

The program is straightforward to use, and readers are invited to use it to explore whatever parameter combinations they are interested in. Because performing a large number of simulations may take considerable processing time (simplicity was favored over speed in the coding), readers who wish to do a large number of simulations are advised to first do a test run using a small number of simulations (e.g., 1000), in order to estimate the processing time per simulation.

Results

FWER

Tables 4, 5, and 6 show the FWERs when population distributions were normal, logistic, and Gumbel, respectively. In each table, the sub-table on the left is for the all-means-equal case, and the sub-table on the right is for the two-means-equal case. $\widehat{\mathrm{SE}} \leq .0002$ for each estimation.

For each procedure, the maximum FWER was .050 for each distribution type. Thus, all procedures controlled the FWER in all examined parameter combinations. FWERs for logistic and Gumbel distributions tended to be lower than corresponding FWERs for normal distributions, but such differences were slight and became increasingly negligible as group size increased (in accordance with the *central limit theorem*).

Table 4

FWERs for Pairwise Comparisons of Three Groups From Normal Distributions

	Population means: {0, 0, 0}					**Population means: {0, 0, 1}**				
	Group size					Group size				
Method	**5**	**10**	**15**	**20**	**1000**	**5**	**10**	**15**	**20**	**1000**
PLSD	.050	.050	.050	.050	.050	.041	.047	.049	.050	.050
HSD	.050	.050	.050	.050	.050	.020	.020	.020	.019	.019
BH	.044	.046	.046	.046	.047	.031	.036	.040	.043	.050

Table 5

FWERs for Pairwise Comparisons of Three Groups From Logistic Distributions

	Population means: {0, 0, 0}					**Population means: {0, 0, 1}**				
	Group size					Group size				
Method	**5**	**10**	**15**	**20**	**1000**	**5**	**10**	**15**	**20**	**1000**
PLSD	.047	.049	.049	.049	.050	.041	.047	.049	.049	.050
HSD	.047	.049	.049	.049	.050	.019	.019	.019	.019	.019
BH	.041	.044	.045	.045	.046	.030	.036	.040	.043	.050

Table 6

FWERs for Pairwise Comparisons of Three Groups From Gumbel Distributions

	Population means: {0, 0, 0}					**Population means: {0, 0, 1}**				
	Group size					Group size				
Method	**5**	**10**	**15**	**20**	**1000**	**5**	**10**	**15**	**20**	**1000**
PLSD	.046	.047	.048	.048	.050	.037	.045	.048	.049	.050
HSD	.045	.047	.048	.048	.050	.019	.019	.019	.019	.019
BH	.040	.042	.044	.044	.046	.028	.034	.038	.042	.050

Note that BH did not exhibit its maximum FWER when all means were equal, but rather when two means were equal and power was maximal (i.e., when group size was very large, though the same high power could have been achieved for smaller group sizes by making μ_3 very large). That is because in the BH algorithm, obtaining a low *p*-value in a given comparison can allow other comparisons to be tested more leniently, meaning that high power in comparisons of groups that truly differ in the population can increase the probability of Type I error in other comparisons. As noted by Finner & Roters (2001), PLSD is somewhat similar to BH in that regard: When a single population mean is different from the rest and power is arbitrarily high, the omnibus test in PLSD is essentially guaranteed to be significant, thereby allowing all pairwise comparisons to be conducted (without adjustment) essentially 100% of the time, thus maximizing the opportunity for Type I error.

Power

Figures 8A and 8B show power versus group-size for the two-means-equal case and all-means-different case, respectively, when population distributions were normal. Results are not shown for group size 1000, because both types of power were always 1 in that case. Per-pair power exhibited a clear hierarchy: PLSD was more powerful than BH (though to an increasingly negligible extent as power increased when all means were different), and BH was more powerful than HSD. Any-pair power also exhibited a clear pattern, though differences between procedures were small: PLSD was marginally more powerful than HSD when two means were equal, and was essentially indistinguishable from HSD when all means were different, whereas BH was the least powerful procedure for both mean configurations by a small margin (and of course the power of all three procedures converged as power approached 1). Altogether, PLSD emerged as the clear winner with regard to power, in that it essentially performed as well or better than the

other procedures in every examined parameter combination and by both definitions of power.

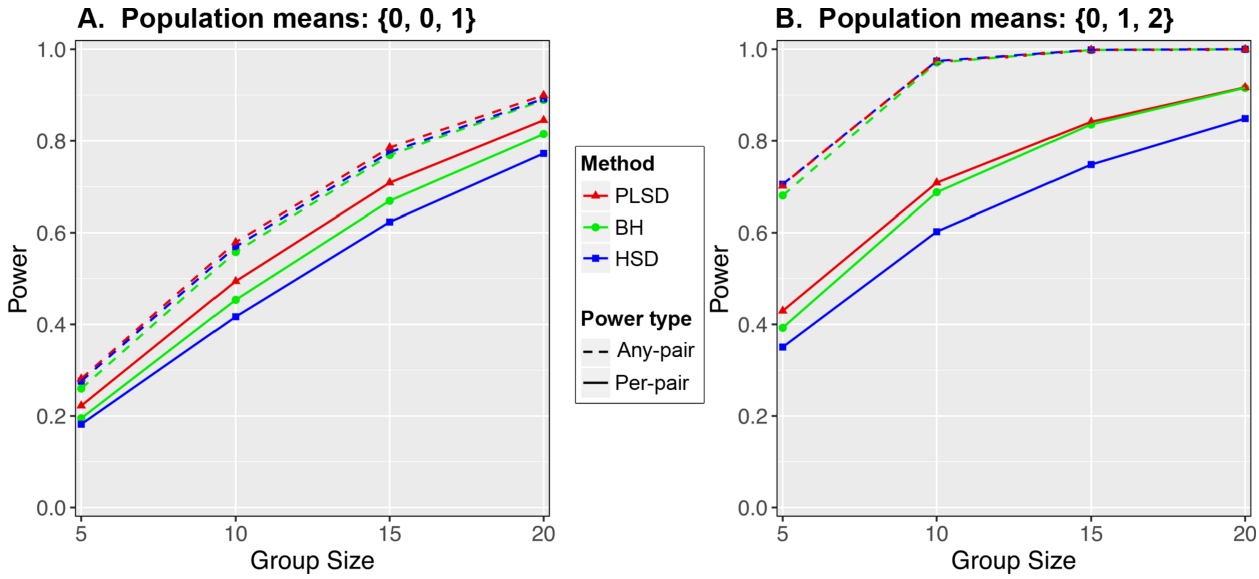

Figure 8. Power of Fisher's protected least significant difference tests (PLSD), Tukey's honest significant difference tests (HSD), and Benjamini–Hochberg adjusted Student *t*-tests (BH), for pairwise comparisons of three groups from normally distributed populations with standard deviation 1. Estimated standard error ≤ .0005 for each point estimation.

Using Gumbel or logistic distributions instead of normal distributions did not alter the power hierarchies described in the preceding paragraph. In fact, power estimates for the Gumbel and logistic distributions were typically only marginally different from the corresponding power estimates under normality. Pooling across all procedures and parameter combinations (excluding group size 1000, for which power was always 1), the mean difference in power between the logistic distribution and the normal distribution was .007 for per-pair power and .004 for any-pair power (favoring the logistic distribution in both cases). And the mean difference in power between the Gumbel distribution and the normal distribution was .012 for per-pair power and .008 for any-pair power (favoring the Gumbel distribution in both cases). These results confirm

that, as previously noted, the dramatically reduced power FM observed for the logistic distribution reflects a methodological inconsistency in their simulations, rather than the non-normality itself. That said, because the logistic and Gumbel distributions are not radically different from the normal distribution in shape, the present results should not be taken to imply that pairwise parametric testing is highly robust to non-normality in general. More substantial departures from normality can have more substantial impact on FWERs and power for parametric tests, particularly when sample sizes are small (Cribbie & Keselman, 2003).

Discussion

For pairwise comparisons of three group means, FM recommended BH-adjusted Student *t*-tests. However, the present study's results suggest that PLSD is a preferable method in that context. Indeed, PLSD was consistently either as powerful as, or more powerful than, both HSD and BH—despite the fact that all three procedures had the same maximum FWER (which was equal to the designated familywise alpha level). This held true for both per-pair power and any-pair power, and held true regardless of whether population distributions were normal, logistic, or Gumbel.

On the other hand, HSD offers a notable feature that PLSD and BH do not: simultaneous confidence intervals. Given that researchers should typically be interested not only in whether means are different, but also in how different the means are, simultaneous confidence intervals are often valuable (Phillips et al., 2013). If using PLSD or BH, one could still report confidence intervals (unadjusted, or perhaps HSD-adjusted), but then the confidence intervals might be incongruent with the significance test results. Thus, heuristically speaking, it appears reasonable in the three-group case to recommend HSD when simultaneous confidence intervals are required,

and to recommend PLSD otherwise (a heuristic that is consistent with Hancock & Klockars, 1996).

A caveat to that heuristic is that both PLSD (in its classical form) and HSD assume that the population distributions have equal variance—an assumption that may often be erroneous (Sauder & DeMars, 2019). The present study did not examine procedures designed for unequal variances (Keselman et al., 1999; Ramsey et al., 2011; Ramsey & Ramsey, 2009; Sauder & DeMars, 2019). However, simulations by Keselman et al. (1999) suggest that in the three-group case, PLSD remains more powerful than BH when adapted for unequal variances (i.e., when using test statistics based on the unpooled standard deviations). Note also that the logic of PLSD remains valid if one substitutes nonparametric tests that do not assume normality (e.g., a Kruskal–Wallis omnibus test followed by Wilcoxon rank-sum tests, rather than an ANOVA followed by *t*-tests).

It is interesting that authors often seem reluctant to recommend PLSD for the three-group case, even if they acknowledge the validity of the approach (e.g., Olejnik & Hess, 1997; Tamhane, 2009, p. 133; Zwick, 1986). Perhaps that reluctance is a reaction to the fact that many misguided researchers have relied on PLSD when comparing more than three groups, under a false sense of security that the FWER was still "protected" from inflation (Keselman et al., 1998; Zwick, 1986). Nonetheless, given that three-group designs are common in experimental research, it is valuable to inform researchers about specialized tools that are optimal for the three-group case (Levin et al., 1994)—even if those tools are not suitable for other cases. If there is concern that recommending valid use of PLSD in the three-group case could inadvertently encourage invalid use of PLSD in other cases, then one could instead recommend a procedure such as Hayter's (1986), which is equivalent to PLSD in the three-group case yet remains valid for larger

numbers of groups (see also Richter & McCann, 2012; Shaffer, 1986). That said, there are often several valid statistical tools to choose from for a given problem, and authors should be cautious about declaring any one approach to be the "best."

EXPERIMENTWISE TYPE I ERROR CONTROL IN 2 × 2 DESIGNS

The present paper proposes three approaches that control the experimentwise Type I error rate in 2 × 2 designs. The first approach, "two-track α_C," tests the interaction at the unadjusted alpha level, and then—using a uniform comparisonwise alpha level determined by simulation—tests either the main effects (if the interaction is nonsignificant) or the simple pairwise effects (if the interaction is significant). The second approach, "one-track α_C," forgoes the main-effect tests and conducts the simple-effect tests using a uniform comparisonwise alpha level determined by simulation. The third approach, "one-track α_H," forgoes the main-effect tests and conducts Hommel-adjusted tests of the simple effects, using a nominal familywise alpha level (determined by simulation) that is marginally higher than the desired experimentwise alpha level. Simulations revealed that the two-track α_C approach tended to retain the most statistical power. However, the other two approaches—because they consistently avoid main-effect tests—may be more straightforwardly interpretable in many circumstances. One-track α_H was more powerful than one-track α_C. But one-track α_C is the only method, of the three methods proposed, that can produce simultaneous confidence intervals.

Introduction

The 2 × 2 design is common in experimental research and can be either between-subjects, within-subjects, or mixed. Tests of *simple effects* (pairwise between-cell differences) in a 2 × 2 design are often, though not always, limited to four meaningful null hypotheses:

$\mu_{11} = \mu_{12}$, $\mu_{11} = \mu_{21}$, $\mu_{12} = \mu_{22}$, and $\mu_{21} = \mu_{22}$, where each μ is the population mean for a given cell, the first subscripted digit denotes the level of Factor A, and the second subscripted digit denotes the level of Factor B. In a 2 × 2 *analysis of variance* (ANOVA), three other null hypotheses are defined: main effect of Factor A ($\mu_{11} + \mu_{12} = \mu_{21} + \mu_{22}$), main effect of Factor B ($\mu_{11} + \mu_{21} = \mu_{12} + \mu_{22}$), and interaction of the two factors ($\mu_{11} - \mu_{12} = \mu_{21} - \mu_{22}$).

If the null hypothesis for the interaction is false, then the effects of the two factors are non-additive and thus cannot be adequately described in terms of main effects. Conversely, if the null hypothesis for the interaction is true, then the null hypotheses for the two simple effects of a given factor are equivalent and thus can be more efficiently tested by collapsing them into a single main-effect test. Therefore, a classical approach in 2 × 2 designs is to test the four simple effects if the interaction is statistically significant, and to test the two main effects if the interaction is nonsignificant (Bethea et al., 1995, p. 233; Bobko, 1986; Devore, 1987, pp. 413–414; Heiberger & Holland, 2004, p. 338; Hochberg & Tamhane, 1987, p. 294; Howell, 2013, p. 420; Nestor & Schutt, 2012, p. 218; Shaffer, 1991; Tamhane, 2009, pp. 232–233; Weber & Skillings, 2000, pp. 429–430). Note that this approach assumes that a significant interaction test would not typically be claimed as a discovery in itself. Rather, in this approach, the interaction test may be considered to function like a railroad track switch that determines whether the analysis should go down the "main-effect track" or the "simple-effect track."

Whether that two-track testing structure is used or not, there is clearly potential for a multiple-comparisons problem in the 2 × 2 design, because there is more than one test of interest. Thus, unless multiple-testing adjustments are made or some *a priori* hierarchical testing structure is established, the *experimentwise Type I error rate* (EWER; the *a priori* probability of at least one Type I error in the results of an experiment) may be inflated. Note that the term *EWER* is

often used interchangeably with the more general term *familywise Type I error rate*, which is the *a priori* probability of at least one Type I error in a "family," i.e., in a set of tests. However, because researchers sometimes divide a single experiment's tests into multiple distinct families—especially in the case of factorial designs—the present paper uses the term *EWER* in order to avoid ambiguity (see Sheskin, 2007, p. 1137). Previously proposed approaches to Type I error control in 2 × 2 designs typically fall into one of two categories: (a) approaches that do not control the EWER, and (b) approaches that can be used to control the EWER but do not accommodate simple-effect testing.

Approaches That Do Not Control The EWER

Often in factorial designs, including 2 × 2 designs, no multiple-testing adjustment whatsoever is performed. In fact, Cramer et al. (2015) reviewed the statistical analyses in six well-known psychology journals and found that although nearly half of the articles used a factorial ANOVA, only 1% of those ANOVA-based analyses addressed Type I error inflation in any way. Moreover, in most articles where a method was used to address Type I error inflation, that method was "omnibus protection"—a method that is known to be unreliable (Kromrey & Dickinson, 1995).

One of the most commonly recommended approaches to Type I error control in 2 × 2 designs is to apply multiple-testing adjustment to the simple-effect tests, using the Bonferroni procedure or some other method, but leave the main-effect tests unadjusted. As Toothaker (1993, p. 79) observed, "most researchers choose to control α for each of the three families of comparisons: the main effect means for A, the main effect means for B, and the cell means" (see also Cardinal & Aitken, 2006; Devore, 1987, pp. 417–418; Girden, 1992, pp. 39–40; Gonzalez, 2009, p. 340; Hancock & Klockars, 1996, p. 276; Hays, 1988, p. 458; Maxwell & Delaney, 2018,

p. 218; Miller, 1981, p. 35; Shaffer, 1995; Tamhane, 2009, pp. 232–233, 242; Weber & Skillings, 2000, pp. 430–431). Taking that approach in the 2 × 2 case means that each main-effect family is comprised of only a single test: level 1 versus level 2 of the given factor. In other words, no multiple-testing adjustment is applied on the main-effect track.

Perhaps the tradition of forgoing adjustment for multiple main-effect tests in the 2 × 2 case stems from the popular belief that tests are exempt from adjustment whenever they are (a) planned *a priori*, (b) orthogonal, and/or (c) few in number. But there appears to be no legitimate basis for any of those arbitrary exemptions, either individually or collectively (as discussed in detail by Frane, 2015b; see also Klockars et al., 1995; Ludbrook, 1991). In fact, orthogonal tests produce *higher* EWERs than positively associated tests (all other things being equal), and the EWER for even just two unadjusted orthogonal tests—whether "planned" or not—is nearly twice the nominal alpha level (when the null hypotheses are true).

Not all authors have endorsed the practice of dividing tests into multiple families by default. Klockars et al. (1995; see also Games, 1971; Hancock & Klockars, 1996) argued that including multiple factors in the same experiment implies that the effects of those factors are being considered jointly as simultaneous inferences, rather than as separate analyses. Similarly, Miller's (1981, p. 34) classic text on multiple comparisons noted that the "natural family" typically consists of the tests for a "single experiment" (see also Ryan, 1962). Miller included two-way ANOVA as an example of a "single experiment," though he acknowledged that there are exceptions and that strict EWER control might not be practical when the number of conditions is large.

Of course, if significance in one of the two main effects in a 2 × 2 design would not constitute a publishable discovery or a basis for decision-making in itself, then not adjusting for

that test would make sense. For instance, that would typically be the case in a study of treatment efficacy where the factor of primary interest was treatment-versus-placebo and the other factor was sex of patient. After all, the goal of the study would be to demonstrate a treatment effect, not to demonstrate a sex effect. It would also likely be inappropriate to adjust the interaction test, because the interaction test would constitute a precaution, rather than an added opportunity to claim "success." However, if the researcher is willing to claim any statistically significant effect as a discovery, as is often be the case in behavioral science research, then there is no apparent justification for exempting any effect tests from adjustment.

In any case, caution should be used when making general statements about how families of tests should be defined. Ultimately, how families should be defined depends on the "problem at hand" (Miller, 1981, p. 35) and on how the test results will be used to make claims or decisions—not on arbitrary criteria such as whether the tests are orthogonal.

Approaches That Do Not Accommodate Simple-Effect Testing

Small et al. (2011) described three methods for controlling the EWER in 2 × 2 designs. Those methods do not predicate simple-effect testing on significance of the interaction. In fact, they do not consider the simple effects at all.

The first method proposed by Small et al. applies the Holm (1979) procedure, which is a stepwise Bonferroni-type adjustment, to a family of three tests: the two main-effect tests and the interaction test. This approach is identical to one later described by Cramer et al. (2015; see their "Remedy 2"; see also Luck & Gaspelin, 2017; Toothaker, 1993, p. 69; Williams, 1973). The second method proposed by Small et al. applies the Holm procedure to the two main-effect tests, and does not consider the interaction. It is not clear why the Holm procedure was recommended for these two approaches, given that other stepwise EWER-controlling procedures (e.g.,

Hommel, 1988) have long been known to preserve more statistical power than the Holm procedure, are valid for most typically-encountered dependence structures, and like the Holm procedure are straightforward to implement using standard software such as R or SAS. The third method proposed by Small et al. is a structured testing approach that consists of the following steps:

1. Test the joint null hypothesis of no main effects at α (the designated experimentwise alpha level). If and only if that test is significant, proceed to Step 2; otherwise, testing stops here and significance of any further tests is forfeited.

2. Test each main effect at α. If and only if both main-effect tests are significant, proceed to Step 3; otherwise, testing stops here and significance of the interaction test is forfeited.

3. Test the interaction at α.

Although the methods proposed by Small et al. (2011; and similarly by Cramer et al., 2015) are mathematically valid, their usefulness is likely to be very limited because they do not accommodate testing of simple effects—even when the interaction appears to be nonzero. That is a nontrivial limitation because stating that an interaction is significant without providing further details is typically not sufficient on its own to support a decision or interesting claim (Bibby, 2012). And although simple-effect testing is not the only way to follow up on a significant interaction (e.g., see Rosnow & Rosenthal, 1989), it is typically the most straightforward and informative way (Howell, 2013, p. 420; Toothaker, 1993, p. 79), especially in the simple 2×2 case. Moreover, the methods proposed by Small et al. require either reducing the power of the interaction test (by subjecting it to Holm adjustment or by predicating it on main-effect

significance) or dispensing with the interaction test entirely.

The Two-Track α_C Method

A simulation-based method is proposed here that controls the EWER when using the aforementioned two-track testing structure. Using simulations to determine optimal multiple-testing adjustments is not a new idea (e.g., see Edwards & Berry, 1987; Hsu & Nelson, 1998; Westfall et al., 2011, pp. 87–90). However, previously established simulation-based methods typically have not accommodated conditional testing structures such as the two-track structure examined here. The presently proposed method consists of the following steps:

1. Simulate data from experiments that have the same design and the same number of observations per cell as the actual experiment, setting all population means equal. In each simulation, the testing is structured as follows: The four simple-effect tests are predicated on interaction significance (at α), and the two main-effect tests are predicated on interaction nonsignificance (at α).

2. Perform a binary search to find a single optimal comparisonwise alpha level (α_C) to use for the simple-effect and main-effect *p*-values that were generated by the tests in Step 1. Specifically, α_C should have a value such that the proportion of simulations that produce at least one significant effect is α; in other words, α_C should produce an estimated EWER equal to α.

3. Using the actual experiment's data, test the interaction at α.

4. If and only if the interaction test was significant, forgo the tests of main effects, and test each of the four simple effects at α_C. If and only if the interaction test was nonsignificant, forgo the tests of simple effects, and test each main effect at α_C.

Appendix D gives an informal proof that this method controls the EWER for all configurations of population means. Note that the method can be reduced to just Steps 3 and 4 by forgoing the simulations and using the appropriate tabulated value of α_C if it is available (see Appendix E for some tabulated values of α_C in balanced designs).

Estimating Two-Track α_C Using R Software

Identifying the optimal two-track α_C is straightforward using R programs that are provided as supplements to the present article: `ac2x2between.R`, `ac2x2within.R`, and `ac2x2mixed.R`, which are for between-subjects, within-subjects, and mixed designs, respectively (see Appendices G, H, and I). Each program performs the simulations and the binary search, and then returns a value of α_C. Without loss of generality, all programs set the population mean to 0 and the population variance to 1 by default.

The code for each program contains a section labeled "INPUT PARAMETERS," which takes several inputs. The first two inputs are the number of simulations (10^7 is recommended) and the α level (presumably .05 in most cases). Next, the user inputs the number of observations in each cell (for between-subjects designs), or the number of subjects (for within-subjects designs), or the number of subjects in each group (for mixed designs). The programs for between-subjects and mixed designs take an additional input: a `TRUE` or `FALSE` logical flag that indicates whether the interaction should be dropped from the model when testing on the main-effect track.

The program for mixed designs also takes an input for population covariance, which is

equivalent to population correlation because the population variance is set to 1. In general, as positive correlation increases, the EWER decreases. Thus, provided that the correlation between repeated measures is assumed to be nonnegative—which is typically a reasonable assumption—a covariance of zero may be used as a "worst case" in the simulations. If negative correlations should be accommodated, then the most negative plausible covariance can be used. It is not necessary to specify population covariance for between-subjects designs, because the observations are independent. And it is not necessary to specify population covariance for within-subjects designs, because population covariance does not affect the EWER of the within-subjects tests when all null hypotheses are true.

Comparison to Two-Track Methods That Do Not Control the EWER

To compare the two-track α_C method to alternative two-track methods, simulations were performed using R programs. Those programs, which are provided as supplements to the present article, are `sim2x2between.R`, `sim2x2within.R`, and `sim2x2mixed.R`, which are for between-subjects, within-subjects, and mixed designs, respectively (see Appendices J, K, and L). The following parameters were used in all simulations: normally distributed populations, equal population-means (specifically equal to 0, without loss of generality), equal population-variance (specifically equal to 1, without loss of generality), and $\alpha = .05$. Various numbers of observations per cell were used. For the within-subjects and mixed designs, the population covariances for the within-subjects variables were set to be equal (specifically equal to 0). Tests on the main-effect track were conducted without the interaction in the model.

Three alternative methods were examined, each of which used the same conditional two-track testing structure as the simulation-based method. The three alternative methods were: (a) Bonferroni adjustment only for explicitly conducted tests, i.e., using a comparisonwise alpha

level of $\alpha/4$ whenever the simple-effect track was reached and $\alpha/2$ whenever the main-effect track was reached, (b) Bonferroni adjustment of the simple-effect tests only, i.e., using a comparisonwise alpha level of $\alpha/4$ whenever the simple-effect track was reached and α whenever the main-effect track was reached, and (c) no adjustment on either track.

Table 7

Estimated Maximum EWERs in 2 × 2 Between-Subjects Designs With Two-Track Approach

# of observations in each cell $\{n_{11}, n_{12}, n_{21}, n_{22}\}$	simulation-based α_C method	Bonferroni adjust only for explicitly conducted tests	Bonferroni adjust simple-effect tests only	no adjustment
{2, 2, 2, 2}	.050	.057	.100	.122
{5, 5, 5, 5}	.050	.065	.110	.132
{20, 20, 20, 20}	.050	.067	.113	.134
{20, 30, 50, 40}	.050	.068	.113	.134
{200, 200, 200, 200}	.050	.068	.113	.135
{2, 20, 2000, 200}	.050	.066	.110	.136

Table 8

Estimated Maximum EWERs in 2 × 2 Within-Subjects Designs With Two-Track Approach

# of subjects	simulation-based α_C method	Bonferroni adjust only for explicitly conducted tests	Bonferroni adjust simple-effect tests only	no adjustment
2	.050	.051	.097	.106
5	.050	.060	.106	.125
10	.050	.065	.111	.132
20	.050	.067	.112	.133
100	.050	.068	.113	.134
1000	.050	.068	.114	.135

The EWER estimates for between-subjects, within-subjects, and mixed designs are shown in Tables 7, 8, and 9, respectively. Each estimate is based on 10^7 simulations. As the tables show, the estimated EWER for the present method was .050 in all cases, whereas the

estimated EWER for each alternative method was inflated in all cases. The fact that EWERs were inflated even when adjusting for all explicitly conducted tests of effects illustrates that the EWER is affected not only by the tests that are formally conducted, but also by other tests that had an *a priori* potential to be conducted before the examination of the data steered the analysis to one track the other. Note also that, for the alternative methods, substituting Holm adjustments for the Bonferroni adjustments would produce the exactly the same maximum EWERs. The estimated standard error ($\widehat{SE}$) of each EWER estimation is $\leq$.0001 (see Albert & Rizzo, 2012, p. 309).

Table 9

Estimated Maximum EWERs in 2 × 2 Mixed Designs With Two-Track Approach

# of subjects in each group	simulation-based α_C method	Bonferroni adjust only for explicitly conducted tests	Bonferroni adjust simple-effect tests only	no adjustment
{2, 2}	.050	.057	.102	.123
{5, 5}	.050	.063	.108	.129
{5, 10}	.050	.064	.110	.132
{20, 20}	.050	.067	.113	.135
{20, 40}	.050	.067	.113	.134
{200, 200}	.050	.068	.114	.135

One-Track Approaches

An inherent problem with the two-track testing structure is that the statistical power of the interaction test is often quite limited and, consequently, analyses often go down the main-effect track even when the interaction is nonzero in the population. This can lead to somewhat meaningless analyses. For example, if a treatment effect substantially differs between male and female populations (i.e., if there is a substantial interaction in the population), then it is

presumably not interesting to test whether there is a treatment effect on average across males and females, because the results would not adequately describe the average person of either sex. Moreover, the main effect could even be driven entirely by an effect in just one of the sexes. In that case, interpreting a significant main effect of treatment as implying a treatment effect in both males and females would be a "de facto Type I error"—even though the main effect itself would not be a Type I error per se. In that scenario, the "de facto EWER" could be much higher than the nominal level, even if the EWER per se were controlled using a procedure such as the two-track α_C method.

Given these concerns, it is arguably better to forgo the main-effect tests altogether in most cases and decide *a priori* to use a "one track" approach that tests only the simple effects. Two such methods are proposed here: the one-track α_C method and the one-track α_H method. Simulations using the `sim2x2between.R`, `sim2x2within.R`, and `sim2x2mixed.R` programs confirmed that both of these methods controlled the EWER at α.

The One-Track α_C Method

The one-track α_C method conducts the simple-effect tests at uniform comparisonwise alpha level α_C, determined by simulation to control the EWER in the given experimental design when all population means are equal (similarly to the two-track α_C method, but without conducting the interaction or main-effect tests). Note that the value of α_C will differ between the one-track and two-track methods. Note also that because the one-track α_C method is an across-the-board alpha adjustment with no conditionality in the testing structure, it can be used to generate simultaneous confidence intervals. This is done by computing each individual confidence interval at the $1 - \alpha_C$ level to achieve experimentwise confidence at the $1 - \alpha$ level.

The One-Track α_H Method

The one-track α_H method is essentially to apply the Hommel (1988) procedure to the simple-effect tests. The Hommel procedure, which retains more statistical power than the Bonferroni and Holm procedures, uses the following algorithm, where α is the experimentwise alpha level, m is the number of tests, $\{p_1, \ldots, p_m\}$ are the p-values ordered from smallest to largest, and b is the integer vector $\{1, \ldots, m\}$:

> Sequentially, for each value of b from 1 to m, if $p_{(m - j + 1)} < (b - j + 1)\ \alpha\ /\ b$ for any $j = \{1, \ldots, b\}$, then $\{p_1, \ldots, p_{(m - b + 1)}\}$ are significant, any other p-values are nonsignificant, and the procedure stops. Else, the procedure continues to the next value of b or, if b has been exhausted, the procedure stops and all p-values are nonsignificant.

The same significance decisions can be obtained by implementing the Hommel procedure as a p-value adjustment (e.g., using the `p.adjust` command in R) and comparing each Hommel-adjusted p-value to α. Note that because the standard Hommel procedure does not take the positive dependence among the tests into account, it can be made slightly more powerful by comparing the adjusted p-values to α_H rather than to α, where α_H is slightly higher than α and is obtained as follows: p-values are simulated for the given experimental design using equal population means (just as in the α_C methods), then Hommel-adjusted p-values are computed, and then a binary search is performed to find the value of α_H that produces the designated EWER.

Values of one-track α_C and α_H can both be obtained using `ac2x2betweensimp.R`, `ac2x2withinsimp.R`, and `ac2x2mixedsimp.R`, which are transcribed in Appendices M, N, and O (respectively) and are analogous to the aforementioned `ac2x2between.R`,

`ac2x2within.R`, and `ac2x2mixed.R` (respectively). Some tabulated values of one-track α_C and α_H are provided in Appendix F. Note that in all cases, α_H is only marginally higher than α. Thus, there appears to be only a marginal power advantage of using the simulation-based value of α_H, rather than simply using α, as the nominal familywise alpha level in the Hommel procedure. Note also that the α_H method allows the maximum EWER to be inflated (up to α_H) when tests are correlated at very near 1, but that is presumably an implausible scenario.

Power Comparisons

Any-Test Power

The `sim2x2between.R`, `sim2x2within.R`, and `sim2x2mixed.R` programs were used to compare the *any-test power* (probability of significance in at least one test of a false null hypothesis) of the three procedures. Tables 10 and 11 show the estimated any-test power for between-subjects designs with 5 and 50 subjects per cell, respectively, using various configurations of standardized population cell-means. Per-test power comparisons for within-subjects and mixed designs showed similar patterns. Each power estimation is based on 10^7 simulations and thus has an $\widehat{SE}$ of $<.0002$.

The two-track α_C method tended to retain more any-test power than the one-track methods, though not universally. That is not surprising, because the main-effect tests have more degrees of freedom than the simple-effect tests, so the inclusion of a main-effect track often offers a power advantage. Indeed, the biggest power advantages for the two-track method occurred when at least one main effect was large relative to the interaction, i.e., when the main-effect track was both likely to be reached and likely to produce significance. Any-test power tended to be slightly higher for the one-track α_H method than for the one-track α_C method.

Table 10

Estimated Any-Test Power in 2 × 2 Between-Subjects Designs With 5 Subjects Per Cell

standardized population cell-means $\{\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}\}$	two-track α_C	one-track α_C	one-track α_H
{0, 0, 0, 2}	.72	.71	.72
{0, 0, 2, 2}	.95	.79	.80
{0, 2, 2, 0}	.88	.87	.90
{0, 1, 1, 3}	.91	.78	.80
{0, 1, 2, 3}	.96	.82	.83

Table 11

Estimated Any-Test Power in 2 × 2 Between-Subjects Designs With 50 Subjects Per Cell

standardized population cell-means $\{\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}\}$	two-track α_C	one-track α_C	one-track α_H
{0, 0, 0, .6}	.81	.84	.84
{0, 0, .5, .5}	.86	.75	.76
{0, .5, .5, 0}	.83	.83	.85
{0, .3, .3, .7}	.77	.62	.64
{0, .3, .4, .7}	.78	.63	.65

Per-Test Power

The `sim2x2between.R`, `sim2x2within.R`, and `sim2x2mixed.R` programs showed that *per-test power* (mean probability of significance across all tests of false null hypotheses) was higher for the one-track α_H method than for the one-track α_C method. Tables 12 and 13 show the estimated per-test power for between-subjects designs with 5 and 50 subjects per cell, respectively, using various configurations of standardized population cell-means. Per-test power comparisons for within-subjects and mixed designs showed similar patterns. Each power estimation is based on 10^7 simulations and thus has an $\widehat{SE}$ of <.0002. There is not a straightforward way to meaningfully compute per-test power for a two-track method, so two-track α_C is not included in these comparisons.

Table 12

Estimated Per-Test Power in 2 × 2 Between-Subjects Designs With 5 Subjects Per Cell

standardized population cell-means $\{\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}\}$	one-track α_C	one-track α_H
{0, 0, 0, 2}	.54	.58
{0, 0, 2, 2}	.54	.58
{0, 2, 2, 0}	.54	.72
{0, 1, 1, 3}	.33	.41

Table 13

Estimated Per-Test Power in 2 × 2 Between-Subjects Designs With 50 Subjects Per Cell

standardized population cell-means $\{\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}\}$	one-track α_C	one-track α_H
{0, 0, 0, .6}	.69	.72
{0, 0, .5, .5}	.50	.53
{0, .5, .5, 0}	.50	.62
{0, .3, .3, .7}	.24	.28

Discussion

Each of the methods proposed here—the two-track α_C method, the one-track α_C method, and the one-track α_H method—reliably controls the EWER. These methods are also straightforwardly modifiable to accommodate different sets of tests than examined here. For instance, if a researcher were interested in five or six simple-effect tests rather than only four, or wished to include the interaction test in the set of adjusted tests, then the simulations used to determine the alpha-level adjustments could be changed accordingly.

In some cases, the two-track α_C method can be considerably more powerful than the one-track α_C and α_H methods. But that power advantage comes at a cost: Whenever the main effect track is taken, any inference about simple effects is sacrificed. That is potentially problematic because main-effect tests often do not directly address the research question of

interest. After all, the very use of the 2 × 2 design in the first place arguably implies that the individual factor levels are relevant. Moreover, a significant main effect may be misinterpreted as implying two simple effects. Thus, the one-track methods may typically be preferable. Of the two one-track methods, the α_H method provides more statistical power, though the one-track α_C method is the only method discussed here that provides valid simultaneous confidence intervals.

To researchers who are accustomed to conducting unadjusted tests in factorial designs, the idea that they should consider substantially reducing their comparisonwise alpha levels in order to control the EWER is likely unwelcome. Nonetheless, those researchers should at least acknowledge that the nominal alpha level they have been using substantially overstates the de facto confidence level of their inferences when the experiment is considered as a whole. Moreover, as the present study's simulations demonstrate, even a "conservative" adjustment such as the Bonferroni procedure allows EWER inflation when some tests are unaccounted for—even if those unaccounted-for tests merely had an *a priori* potential to be conducted and were not formally performed in the analysis.

The present paper does not take the view that strict EWER control is mandatory in all cases. In fact, statisticians have long recognized that the proper way to define the family for multiple-testing purposes depends on the research situation (e.g., Miller, 1981, pp. 31–35). However, controlling the EWER level is presumably desirable in many studies, including some studies that use the common 2 × 2 design (Hancock & Klockars, 1996; Klockars et al., 1995; Miller, 1981, pp. 34–35). And although defining families of tests at some "sub-experimentwise" level may sometimes be reasonable, that decision should not be made thoughtlessly.

APPENDIX A:

DEMONSTRATING THAT MANOVA-PROTECTION WITH THE α / (m – 1) ADJUSTMENT CONTROLS THE PER-FAMILY TYPE I ERROR RATE

When There Is an Effect in One or More Outcome Variables

MANOVA protection (MP, as defined on pp. 52–53 of this dissertation) allows univariate testing only if the omnibus test is significant. Thus, for the purpose of computing MP's maximum per-family Type I error rate (PFER) when there is at least one nonzero effect in the population, it is sufficient to assume the "worst case" scenario that the omnibus test is always significant (as would be the case if at least one effect were arbitrarily large) and simply consider the univariate tests.

In a two-group multivariate design, if there is a nonzero effect in one outcome variable, then there are at most $m - 1$ opportunities for Type I error among the m univariate tests. For unadjusted alpha level α, this makes the maximum PFER equal to $\alpha \times (m - 1)$. However, substituting $[\alpha / (m - 1)]$ for α as the testwise alpha level for the univariate tests (as described on pp. 52–53 of this dissertation) caps the PFER at $[\alpha / (m - 1)] \times (m - 1) = \alpha$.

If there are nonzero effects in more than one outcome variable, then the maximum PFER is even lower. Indeed, it is clear that $[\alpha / (m - 1)] \times (m - m_1) < \alpha$ for any $m_1 > 1$, where m_1 is the number of nonzero effects.

When There Is No Effect in Any Outcome Variable

When all effect sizes are zero, population correlations among outcome variables do not affect MP's PFER, as can be straightforwardly confirmed by simulation. It is therefore sufficient to demonstrate control of the PFER when all population correlations are zero.

An original R (version 3.0.2; R Core Team, 2013) program was used to estimate the PFER of MP with the $\alpha / (m - 1)$ adjustment, with all effect sizes and all population correlations set to zero. Various values of n (sample size per group) and α were used. The univariate tests were two-sided two-sample t-tests using the pooled standard deviation.

For $\alpha = .05$, $n = 50$, observed PFERs were as follows for each value of m from 2 to 10, respectively: 0.049, 0.041, 0.036, 0.033, 0.030, 0.028, 0.026, 0.025, and 0.024 (estimated standard error ≤ 0.0002 for all estimates, based on 1,000,000 simulations for each value of m). Simulations using $n = 30$, $n = 500$, and $n = 5{,}000$ produced similar results. Simulations using very small n controlled the PFER more conservatively.

Simulations using $\alpha < .05$ (e.g., $\alpha = .01$) consistently produced PFERs that were lower than the given α. But some simulations using $\alpha > .05$ allowed PFER inflation; this is because the occurrence of Type I errors in multiple tests at once—which is a relatively rare occurrence when α is low—becomes increasingly common as α increases, thus expanding the disparity between PFER and FWER.

In summary, MP with the $\alpha / (m - 1)$ adjustment appears to reliably control the PFER for any number of outcome variables, unless an unconventionally high value of α ($>.05$) is used. In fact, when all null hypotheses are true, MP with the $\alpha / (m - 1)$ adjustment controls the PFER with increasing conservatism as m increases.

APPENDIX B:

MANOVA's "WEAK SPOTS"

When there are two outcome variables, the formula for the multivariate analysis of variance (MANOVA) noncentrality parameter is as follows (adapted from Morrison, 1967):

$$\delta^2 = \frac{n}{2} \times \frac{\Delta_X{}^2 - 2\rho\Delta_X\Delta_Y + \Delta_Y{}^2}{1-\rho^2},$$

where n is the sample size per group, Δ_X and Δ_Y are the standardized effect sizes in Variables X and Y respectively, and ρ is the population correlation between X and Y. MANOVA's statistical power is a monotonically increasing function of noncentrality parameter δ^2. Given values of n and any two of the three parameters (Δ_X, Δ_Y, ρ), δ^2 has local minimums ("weak spots") where $\Delta_X = \rho\Delta_Y$ or where $\Delta_Y = \rho\Delta_X$, as can be demonstrated with elementary calculus. It follows algebraically that MANOVA's weak spots occur where ρ is equal to the ratio of the smaller effect size to the larger effect size. What may be less obvious is that these weak spots generalize beyond the two-variable case to any number of outcome variables as follows:

> Let Variable_1 be the outcome variable with the largest effect size in absolute value. Then MANOVA's power is minimized where each bivariate correlation that includes Variable_1 is equal to the effect size ratio of those two variables. For example, if there are 3 outcome variables (X, Y, and Z) and Δ_Z is the largest effect size in absolute value, then MANOVA's weak spots occur where $\rho_{XZ} = \Delta_X / \Delta_Z$ and $\rho_{YZ} = \Delta_Y / \Delta_Z$ (subscripts of ρ indicating the variables in the given bivariate correlation). At these weak spots, δ^2 reduces to $(n / 2)\,\Delta_Z{}^2$. Thus, for any given sample size, MANOVA's power at the weak

spot is simply a function of Δ_Z^2. The theorem in the next section formalizes these properties of MANOVA's weak spots for any number of variables.

MANOVA Weak-Spot Theorem

Let **A** be the correlation matrix for $\text{Variable}_1, \ldots, \text{Variable}_m$. Let $\mathbf{\Delta} = (\Delta_1, \ldots, \Delta_m)^T$ be the column vector of standardized effect sizes for m variables for the two-group Hotelling's T^2 distribution. Without loss of generality, let $|\Delta_1| > |\Delta_i|$ for all $i > 1$. Let the elements of **A** be denoted as ρ_{ij}. Then:

1. For the quadratic form $Q = \mathbf{\Delta}^\mathbf{T}\mathbf{A}^{-1}\mathbf{\Delta}$, which is used in computing the MANOVA noncentrality parameter $\delta^2 = (n / 2) \times Q$, where $Q \geq \Delta_1^2$.

2. For a given value of Δ_1 and given values of ρ_{i1} for $i = 2, \ldots, m$ (i.e., the correlations of Variable_1 with each of the other variables), the minimum value Δ_1^2 for Q is attained when $\Delta_i = \rho_{i1}\Delta_1$ independently of the correlations ρ_{ij} (where $j > 1$). That is, it does not matter what these other correlations are as long as **A** is a nonsingular correlation matrix.

3. For a given correlation matrix **A** and given the effect size vector $\mathbf{\Delta}$ such that $|\Delta_1| > |\Delta_i|$ for all $i > 1$, Q is minimized and is equal to Δ_1^2 when $\rho_{i1} = \Delta_i / \Delta_1$, with no constraints on the other correlations other than that **A** be a nonsingular correlation matrix.

A proof for the above theorem, using correspondingly numbered steps, is as follows:

1. Let $\text{Variable}_2, \ldots, \text{Variable}_m$ be transformed to be orthogonal to Variable_1. This is done by replacing Variable_i by $(\text{Variable}_i - \rho_{1i} \times \text{Variable}_1)$ for all $i > 1$, where ρ_{1i} is the correlation between Variable_1 and Variable_i. Then each transformed variable is rescaled

to have unit variance by dividing by $(1 - \rho_{1i}^2)$. Let **C** be the correlation matrix for the resulting transformed variables and let **Δ*** be the vector of the corresponding transformed standardized effect sizes.

The value of Q is unchanged by linear, non-singular transformations of the variables. Hence, $Q = \mathbf{\Delta}^T\mathbf{A}^{-1}\mathbf{\Delta} = \mathbf{\Delta}^{*T}\mathbf{C}^{-1}\mathbf{\Delta}^*$, where $\mathbf{\Delta}^* = (\Delta_1, \Delta_2^*, \ldots, \Delta_m^*)^T$. The first row and first column elements of **C** are $c_{1i} = 0$ and $c_{i1} = 0$ (for all $i > 1$) because the transformed variables have been made to be uncorrelated with Variable_1. This implies that the first row and first column elements of $\mathbf{C}^{-1}$ (i.e., **C***) are $c^*_{11} = 1$, $c^*_{1i} = 0$, and $c^*_{i1} = 0$ (for all $i > 1$).

Let $\mathbf{C}_S$ denote the submatrix of **C** excluding the first row and first column. Then $Q = \Delta_1^2 + (\Delta_2^*, \ldots, \Delta_m^*)^T\mathbf{C}_S^{-1}(\Delta_2^*, \ldots, \Delta_m^*)$, where the second term (i.e., the quadratic form involving only transformed $\text{Variable}_2, \ldots, \text{Variable}_m$) is greater than or equal to zero. Hence, $Q \geq \Delta_1^2$.

2. Let $\text{Variable}_{m+1} = \text{Variable}_1 \times \Delta_1$. The least squares solution for the regression of Variable_{m+1} on Variable_1 to Variable_m is simply $\text{Variable}_{m+1} = \text{Variable}_1 \times \Delta_1$. Therefore, the solution vector is $(\Delta_1, 0, \ldots, 0)^T$. The least squares solution is also equal to $\mathbf{A}^{-1}(\Delta_1, \ldots, \Delta_m)^T = \mathbf{A}^{-1}(\Delta_1, \rho_{21}\Delta_1, \ldots, \rho_{m1}\Delta_1)^T$, and so $\mathbf{\Delta}^T\mathbf{A}^{-1}\mathbf{\Delta} = \mathbf{\Delta}^T(\Delta_1, 0, \ldots, 0)^T = \Delta_1^2$.

3. It is also clear that if correlations are so defined from the effect sizes, then $Q = \mathbf{\Delta}^T\mathbf{A}^{-1}\mathbf{\Delta} = \Delta_1^2$.

APPENDIX C:

R CODE FOR SIMULATING PAIRWISE COMPARISONS

OF THREE GROUPS

```
# This program performs simulations to compute the power and familywise Type I error
# rates (FWERs) of Benjamini-Hochberg ("BH"; applied to Student t-tests), Tukey's HSD,
# and Fisher's protected LSD ("PLSD"), for all possible group-mean comparisons in a
# balanced 3-group design where equal variance is assumed.

# "Average power" is per-pair power, i.e., the mean power across all tests of false
# null hypotheses. "Any-pair power" is the power to get significance in any 1 or more
# tests of false null hypotheses.

# Author: Andrew V. Frane
# Created in 2018 (and revised in 2019) using R version 3.3.3.

rm(list=ls())              # clear variables from workspace
startTime = proc.time()[3] # start timing this program

################################################################################

# ADJUSTABLE PARAMETERS

# Notes: Setting all standard deviations to 1 allows means to be considered as
# standardized. Unequal standard deviations can be entered, but the tests will still
# assume equal variance. Similarly, non-normal distribution can be entered, but the
# tests will still be parametric.

numSim    = 10^6       # number of simulations
groupSize = 5          # sample size per group

mu1       = 0          # population mean for group 1 (can be any real value)
mu2       = 0          # population mean for group 2 (can be any real value)
mu3       = 0          # population mean for group 3 (can be any real value)

sd1       = 1          # population standard deviation for group 1
sd2       = 1          # population standard deviation for group 2
sd3       = 1          # population standard deviation for group 3

a         = .05        # familywise alpha level

popDist   = "normal"   # population distribution type
                       # (can be "normal", "logistic", or "Gumbel")

################################################################################

# INITIALIZE P-VALUE VECTORS

pOmni      = vector("double", numSim) # ANOVA omnibus p-values

p1vs2TTest = vector("double", numSim) # raw t-test     p-values for group 1 vs group 2
p1vs3TTest = vector("double", numSim) # raw t-test     p-values for group 1 vs group 3
p2vs3TTest = vector("double", numSim) # raw t-test     p-values for group 2 vs group 3
```

```
p1vs2BH    = vector("double", numSim) # BH-adjusted    p-values for group 1 vs group 2
p1vs3BH    = vector("double", numSim) # BH-adjusted    p-values for group 1 vs group 3
p2vs3BH    = vector("double", numSim) # BH-adjusted    p-values for group 2 vs group 3

p1vs2Tukey = vector("double", numSim) # Tukey-adjusted p-values for group 1 vs group 2
p1vs3Tukey = vector("double", numSim) # Tukey-adjusted p-values for group 1 vs group 3
p2vs3Tukey = vector("double", numSim) # Tukey-adjusted p-values for group 2 vs group 3

p1vs2PLSD  = vector("double", numSim) # PLSD-adjusted  p-values for group 1 vs group 2
p1vs3PLSD  = vector("double", numSim) # PLSD-adjusted  p-values for group 1 vs group 3
p2vs3PLSD  = vector("double", numSim) # PLSD-adjusted  p-values for group 2 vs group 3

################################################################################

# GENERATE "MASTER VECTORS" OF OBSERVATIONS FROM DESIRED RANDOM DISTRIBUTION

# Each simulation will use a different segment of each master vector.
# Note that in the Gumbel location parameter, digamma(1) is -1 times Euler's constant.

yMasterLength = numSim*groupSize # length of each master vector

if (popDist == "normal" | popDist == "Normal") {
  yMaster1 = rnorm(yMasterLength, mu1, sd1) # master vector for group 1
  yMaster2 = rnorm(yMasterLength, mu2, sd2) # master vector for group 2
  yMaster3 = rnorm(yMasterLength, mu3, sd3) # master vector for group 3

} else if (popDist == "logistic" | popDist == "Logistic") {
  yMaster1 = rlogis(yMasterLength, mu1, sd1*sqrt(3)/pi) # master vector for group 1
  yMaster2 = rlogis(yMasterLength, mu2, sd2*sqrt(3)/pi) # master vector for group 2
  yMaster3 = rlogis(yMasterLength, mu3, sd3*sqrt(3)/pi) # master vector for group 3

} else if (popDist == "gumbel" | popDist == "Gumbel") {

  if(!require(evd)) {       # try to load 'evd' package; and if it isn't found...
    install.packages("evd") # ...install it...
    require(evd)            # ...and load it
  }

  kScale = sqrt(6)/pi # constant used in scale parameters

  yMaster1 = rgumbel(yMasterLength, mu1 + sd1*kScale*digamma(1), sd1*kScale) # group 1
  yMaster2 = rgumbel(yMasterLength, mu2 + sd2*kScale*digamma(1), sd2*kScale) # group 2
  yMaster3 = rgumbel(yMasterLength, mu3 + sd3*kScale*digamma(1), sd3*kScale) # group 3

} else { # force error and display warning if valid distribution type not assigned
  yMaster1 = NA
  yMaster2 = NA
  yMaster3 = NA

  warning(popDist, " is not a valid input for population distribution type")
  }

# initialize index-range defining where in master vectors a given simulation's
# observations come from (first number is start index; second number is end index)
yIndexRange = c(1, groupSize)

################################################################################

# SIMULATIONS (COMPUTATION OF P-VALUES)

for (iSim in 1:numSim) {
  y1 = yMaster1[yIndexRange[1] : yIndexRange[2]] # grp 1 observations for current sim
  y2 = yMaster2[yIndexRange[1] : yIndexRange[2]] # grp 2 observations for current sim
```

```
  y3 = yMaster3[yIndexRange[1] : yIndexRange[2]] # grp 3 observations for current sim

  # ANOVA ('dframe' is the data matrix)
  dframe      = data.frame(grp=rep(c("1", "2", "3"), each=groupSize), y=c(y1, y2, y3))
  anovaModel  = aov(y ~ grp, dframe)                          # ANOVA model
  pOmni[iSim] = summary(anovaModel) [[1]]["grp", "Pr(>F)"] # omnibus p-value

  # Student t-tests using pooled standard deviation from all three groups
  tTestTable = pairwise.t.test(dframe$y, dframe$grp, p.adjust.method="none",
                               pool.sd=TRUE) # student t-test table

  p1vs2TTest[iSim] = tTestTable$p.value["2", "1"] # raw p-value for t-test grp 1 vs. 2
  p1vs3TTest[iSim] = tTestTable$p.value["3", "1"] # raw p-value for t-test grp 1 vs. 3
  p2vs3TTest[iSim] = tTestTable$p.value["3", "2"] # raw p-value for t-test grp 2 vs. 3

  # Benjamini-Hochberg ('pBH' is the vector of BH-adjusted p-values for current sim)
  pBH = p.adjust(c(p1vs2TTest[iSim], p1vs3TTest[iSim], p2vs3TTest[iSim]), "BH")

  p1vs2BH[iSim] = pBH[1] # BH-adjusted p-value for group 1 vs. group 2
  p1vs3BH[iSim] = pBH[2] # BH-adjusted p-value for group 1 vs. group 3
  p2vs3BH[iSim] = pBH[3] # BH-adjusted p-value for group 2 vs. group 3

  # Tukey's HSD
  tukeyTable = TukeyHSD(anovaModel) # table of Tukey's HSD results

  p1vs2Tukey[iSim] = tukeyTable$grp["2-1", "p adj"] # HSD-adjusted p-value grp 1 vs. 2
  p1vs3Tukey[iSim] = tukeyTable$grp["3-1", "p adj"] # HSD-adjusted p-value grp 1 vs. 3
  p2vs3Tukey[iSim] = tukeyTable$grp["3-2", "p adj"] # HSD-adjusted p-value grp 2 vs. 3

  yIndexRange = yIndexRange + groupSize # update index-range for next simulation
}

# Fisher's protected LSD
p1vs2PLSD = pmax(p1vs2TTest, pOmni) # PLSD-adjusted p-value for group 1 vs. group 2
p1vs3PLSD = pmax(p1vs3TTest, pOmni) # PLSD-adjusted p-value for group 1 vs. group 3
p2vs3PLSD = pmax(p2vs3TTest, pOmni) # PLSD-adjusted p-value for group 2 vs. group 3

################################################################################

# ESTIMATE FAMILYWISE ERROR RATE (FWER), AVG POWER, & ANY-PAIR POWER FOR EA. PROCEDURE

# FWER
fwerBH    = mean((mu1 == mu2 & p1vs2BH <= a) | (mu1 == mu3 & p1vs3BH <= a) |
                 (mu2 == mu3 & p2vs3BH <= a)) # FWER for BH

fwerTukey = mean((mu1 == mu2 & p1vs2Tukey <= a) | (mu1 == mu3 & p1vs3Tukey <= a) |
                 (mu2 == mu3 & p2vs3Tukey <= a)) # FWER for Tukey's HSD

fwerPLSD  = mean((mu1 == mu2 & p1vs2PLSD  <= a) | (mu1 == mu3 & p1vs3PLSD  <= a) |
                 (mu2 == mu3 & p2vs3PLSD  <= a)) # FWER for PLSD

# any-pair power
anyPowerBH    = mean((mu1 != mu2 & p1vs2BH <= a) | (mu1 != mu3 & p1vs3BH <= a) |
                     (mu2 != mu3 & p2vs3BH <= a)) # any-pair power for BH

anyPowerTukey = mean((mu1 != mu2 & p1vs2Tukey <= a) | (mu1 != mu3 & p1vs3Tukey <= a) |
                     (mu2 != mu3 & p2vs3Tukey <= a)) # any-pair power for Tukey's HSD

anyPowerPLSD  = mean((mu1 != mu2 & p1vs2PLSD  <= a) | (mu1 != mu3 & p1vs3PLSD  <= a) |
                     (mu2 != mu3 & p2vs3PLSD  <= a)) # any-pair power for PLSD

# average (per-pair) power
numFalseNulls = sum(c(mu1 != mu2, mu1 != mu3, mu2 != mu3)) # number of false nulls
```

```
avgPowerBH    = mean(((mu1 != mu2 & p1vs2BH <= a) + (mu1 != mu3 & p1vs3BH <= a) +
                      (mu2 != mu3 & p2vs3BH <= a)) / numFalseNulls) # avg power for BH

avgPowerTukey = mean(((mu1 != mu2 & p1vs2Tukey <= a) + (mu1 != mu3 & p1vs3Tukey <= a)+
                      (mu2 != mu3 & p2vs3Tukey <= a))/ numFalseNulls) # avg power HSD

avgPowerPLSD  = mean(((mu1 != mu2 & p1vs2PLSD  <= a) + (mu1 != mu3 & p1vs3PLSD  <= a)+
                      (mu2 != mu3 & p2vs3PLSD  <= a))/ numFalseNulls) # avg power PLSD

################################################################################

# ESTIMATED STANDARD-ERRORS FOR THE ABOVE ESTIMATIONS

# Estimated SE = sqrt(v*(1-v)/numSim), where v is the FWER estimate or power estimate

fwerBHSE    = sqrt(fwerBH   *(1-fwerBH   )/numSim) # estimated SE for FWER of BH
fwerTukeySE = sqrt(fwerTukey*(1-fwerTukey)/numSim) # estimated SE for FWER of HSD
fwerPLSDSE  = sqrt(fwerPLSD *(1-fwerPLSD )/numSim) # estimated SE for FWER of PLSD

avgPowerBHSE    = sqrt(avgPowerBH   *(1-avgPowerBH   )/numSim) # estimSE avgPower BH
avgPowerTukeySE = sqrt(avgPowerTukey*(1-avgPowerTukey)/numSim) # estimSE avgPower HSD
avgPowerPLSDSE  = sqrt(avgPowerPLSD *(1-avgPowerPLSD )/numSim) # estimSE avgPower PLSD

anyPowerBHSE    = sqrt(anyPowerBH   *(1-anyPowerBH   )/numSim) # estimSE anyPower BH
anyPowerTukeySE = sqrt(anyPowerTukey*(1-anyPowerTukey)/numSim) # estimSE anyPower HSD
anyPowerPLSDSE  = sqrt(anyPowerPLSD *(1-anyPowerPLSD )/numSim) # estimSE anyPower PLSD

fwerMaxSE     = max(fwerBHSE    , fwerTukeySE    , fwerPLSDSE    ) # maxEstimSE FWER
avgPowerMaxSE = max(avgPowerBHSE, avgPowerTukeySE, avgPowerPLSDSE) # maxEstimSE avgPwr
anyPowerMaxSE = max(anyPowerBHSE, anyPowerTukeySE, anyPowerPLSDSE) # maxEstimSE anyPwr

################################################################################

# REPORT THE PARAMETERS AND RESULTS

numSim           # number of simulations
a                # designated familywise alpha level
fwerMaxSE        # maximum estimated standard error for FWER
avgPowerMaxSE    # maximum estimated standard error for average power
anyPowerMaxSE    # maximum estimated standard error for any-pair power
popDist          # population distribution type
groupSize        # sample size per group
c(mu1, mu2, mu3) # array of population means
c(sd1, sd2, sd3) # array of population standard deviations

fwerBH           # FWER for Benjamini-Hochberg
fwerTukey        # FWER for Tukey's HSD
fwerPLSD         # FWER for PLSD
avgPowerBH       # average  power for Benjamini-Hochberg
avgPowerTukey    # average  power for Tukey
avgPowerPLSD     # average  power for PLSD
anyPowerBH       # any-pair power for Benjamini-Hochberg
anyPowerTukey    # any-pair power for Tukey's HSD
anyPowerPLSD     # any-pair power for PLSD

(proc.time() [3] - startTime) / 60 # total time elapsed in minutes
```

APPENDIX D:

PROOF THAT THE TWO-TRACK SIMULATION-BASED α_C METHOD CONTROLS THE EXPERIMENTWISE TYPE I ERROR RATE IN 2×2 DESIGNS

Before constructing the proof that the two-track simulation-based α_C method (as described on pp. 104–105 in this dissertation) controls the experimentwise Type I error rate (EWER), it is helpful to establish two premises:

> **Premise 1**. Because $\alpha_C < \alpha / 2$ in all cases (see Appendix E), α_C controls the EWER at $<\alpha$ whenever the number of true null hypotheses is ≤ 2 (by the principles of the Bonferroni procedure; Dunn, 1961).
>
> **Premise 2.** By elementary principles of probability, under the defined testing structure, EWER = P(significant interaction) × P(at least one simple-effect Type I error | significant interaction) + P(nonsignificant interaction) × P(at least one main-effect Type I error | nonsignificant interaction).

The proof now follows for each possible configuration of population means.

When all means are equal. In this scenario, the method is essentially self-validating by virtue of how α_C is obtained. That is, α_C is demonstrated by simulation to control the EWER when all population means are equal.

When three means are equal. Without loss of generality, let $\mu_{11} = \mu_{12} = \mu_{21} \neq \mu_{22}$. In

this scenario, there are only two true null hypotheses for the simple-effects: $\mu_{11} = \mu_{12}$ and $\mu_{11} = \mu_{21}$. And there are no true null hypotheses for the main effects. Thus, because there are only two true null hypotheses of interest, the EWER is controlled as per Premise 1.

When two adjacent means are equal. The term *adjacent means* is used here to refer to means that share a factor level and are thus in the same row or column of the 2 × 2 design. Without loss of generality, let $\mu_{11} = \mu_{12}$, and let all other mean pairings be unequal. In this scenario, there is only one true null hypothesis for the simple-effect tests: $\mu_{11} = \mu_{12}$. And there can be at most one true null hypothesis for the main effect tests: $\mu_{11} + \mu_{12} = \mu_{21} + \mu_{22}$. Thus, there are at most only two true null hypotheses of interest, so the EWER is controlled as per Premise 1.

When two parallel pairs of adjacent means are equal. Without loss of generality, let $\mu_{11} = \mu_{12} \neq \mu_{21} = \mu_{22}$. In this scenario, there are three true null hypotheses: two for the simple effects ($\mu_{11} = \mu_{12}$ and $\mu_{21} = \mu_{22}$), and one for a main effect ($\mu_{11} + \mu_{21} = \mu_{12} + \mu_{22}$.). As in the all-means-equal scenario, the null hypothesis for the interaction is true, so P(significant interaction) and P(nonsignificant interaction) in the present scenario are the same as in the all-means-equal scenario; specifically, P(significant interaction) = α, and P(nonsignificant interaction) = $1 - \alpha$. Moreover, the distributions of the test statistics for the tests of true null hypotheses in the current scenario are the same as in the all means-true-scenario—there are just fewer of them. Thus, neither P(at least one simple-effect Type I error | significant interaction) nor P(at least one main-effect Type I error | nonsignificant interaction) can be higher in the present scenario than in the all-means-equal scenario. In other words, in the present scenario, all the probabilities in the Premise 2 formula are less than or equal to what they would be in the all-means-equal scenario. Consequently, any method that controls the EWER in the all-means-equal scenario controls the EWER conservatively in the present scenario.

When two nonadjacent means are equal. Without loss of generality, let $\mu_{11} = \mu_{22}$, and let all other mean pairings be unequal. In this scenario, there are no true null hypotheses for the simple-effect or main-effect tests of interest, so there cannot be any Type I errors.

When two pairs of nonadjacent means are equal. In this scenario ($\mu_{11} = \mu_{22} \neq \mu_{12} = \mu_{21}$), the null hypotheses for both main effects are true, but there are no true null hypotheses for the simple effects. Thus, there are only two true null hypotheses so the EWER is controlled as per Premise 1.

When all means are different. In this case, there are no true null hypotheses for the simple effects, and there can be at most one true null hypothesis for the main effects. Because there cannot be more than one true null hypothesis, testing at any level less than or equal to α controls the EWER.

APPENDIX E:

TABULATED VALUES OF TWO-TRACK α_C

FOR BALANCED 2 × 2 DESIGNS

Tables 14, 15, and 16 give values of α_C (as defined on pp. 104–105 of this dissertation) for between-subjects, within-subjects, and mixed 2 × 2 designs, respectively, using $\alpha = .05$ and equal numbers of observations per cell. Each estimate of α_C is based on 3×10^7 simulations. Note that in all cases, $\alpha_C < \alpha / 2$, and the highest value of α_C is obtained when the sample size is minimal; additional simulations (not detailed here) confirmed that the same is true for less conventional values of α, such as .001, .005, .01, and .1.

Table 14

Values of Two-Track α_C for Balanced Between-Subjects Designs (α = .05)

observations per cell	α_C	α_C as a function of α
2	.0184	α / 2.72
3	.0164	α / 3.06
4	.0156	α / 3.20
5	.0153	α / 3.28
6	.0151	α / 3.31
7	.0149	α / 3.35
8	.0149	α / 3.37
9	.0148	α / 3.38
10	.0147	α / 3.39
11	.0147	α / 3.40
12	.0147	α / 3.41
13	.0146	α / 3.41
14	.0146	α / 3.42
15	.0146	α / 3.43
16	.0146	α / 3.43
17	.0146	α / 3.43
18	.0145	α / 3.44
19	.0145	α / 3.44
20	.0145	α / 3.44
30	.0145	α / 3.46
40	.0144	α / 3.46
50	.0144	α / 3.47
100	.0144	α / 3.47
1,000	.0144	α / 3.48
1,000,000,000	.0144	α / 3.48

Note. In the underlying simulations, main-effect tests were conducted without the interaction in the model.

Table 15

Values of Two-Track α_C for Within-Subjects Designs (α = .05)

number of subjects	α_C	α_C as a function of α
2	.0229	α / 2.18
3	.0203	α / 2.46
4	.0184	α / 2.71
5	.0174	α / 2.88
6	.0167	α / 3.00
7	.0162	α / 3.08
8	.0159	α / 3.14
9	.0157	α / 3.18
10	.0155	α / 3.22
11	.0154	α / 3.25
12	.0153	α / 3.27
13	.0152	α / 3.29
14	.0151	α / 3.30
15	.0151	α / 3.32
16	.0150	α / 3.33
17	.0150	α / 3.34
18	.0149	α / 3.35
19	.0148	α / 3.37
20	.0149	α / 3.37
30	.0147	α / 3.40
40	.0146	α / 3.42
50	.0146	α / 3.43
100	.0144	α / 3.47
1,000	.0144	α / 3.48
1,000,000,000	.0144	α / 3.48

Table 16

Values of Two-Track α_C for Balanced Mixed Designs (α = .05)

subjects per group	α_C	α_C as a function of α
2	.0190	α / 2.63
3	.0170	α / 2.94
4	.0161	α / 3.10
5	.0156	α / 3.20
6	.0153	α / 3.26
7	.0151	α / 3.30
8	.0150	α / 3.33
9	.0149	α / 3.35
10	.0149	α / 3.36
11	.0148	α / 3.38
12	.0148	α / 3.39
13	.0147	α / 3.40
14	.0147	α / 3.41
15	.0147	α / 3.41
16	.0146	α / 3.42
17	.0146	α / 3.42
18	.0146	α / 3.43
19	.0146	α / 3.43
20	.0146	α / 3.43
30	.0145	α / 3.44
40	.0144	α / 3.46
50	.0144	α / 3.47
100	.0144	α / 3.47
1,000	.0144	α / 3.48
1,000,000,000	.0144	α / 3.48

Notes. The underlying simulations used a population covariance of zero (i.e., nonnegative covariance was assumed). Main-effect tests were conducted without the interaction in the model.

APPENDIX F:

TABULATED VALUES OF ONE-TRACK α_C AND α_H

FOR BALANCED 2 × 2 DESIGNS

Tables 17, 18, and 19 give values of one-track α_C and α_H (as defined on pp. 109–110 of this dissertation) for between-subjects, within-subjects, and mixed 2 × 2 designs, respectively, using $\alpha = .05$ and equal numbers of observations per cell. The estimates in each row of each table are based on 10^7 simulations.

Table 17

Values of One-Track α_C and α_H for Balanced Between-Subjects Designs ($\alpha = .05$)

observations per cell	α_C	α_C as a function of α	α_H
2	.0138	α / 3.62	.0532
3	.0139	α / 3.60	.0536
4	.0140	α / 3.57	.0540
5	.0140	α / 3.57	.0541
6	.0141	α / 3.58	.0542
7	.0141	α / 3.55	.0543
8	.0141	α / 3.55	.0543
9	.0141	α / 3.54	.0545
10	.0141	α / 3.54	.0545
100	.0141	α / 3.54	.0545
1,000	.0141	α / 3.54	.0545
1,000,000,000	.0141	α / 3.54	.0546

Table 18

Values of One-Track α_C and α_H for Within-Subjects Designs (α = .05)

observations per cell	α_C	α_C as a function of α	α_H
2	.0128	α / 3.92	.0504
3	.0129	α / 3.88	.0506
4	.0130	α / 3.85	.0509
5	.0132	α / 3.79	.0515
6	.0133	α / 3.76	.0517
7	.0134	α / 3.73	.0521
8	.0135	α / 3.71	.0524
9	.0136	α / 3.69	.0526
10	.0136	α / 3.68	.0527
100	.0141	α / 3.55	.0544
1,000	.0141	α / 3.55	.0544
1,000,000,000	.0141	α / 3.55	.0544

Table 19

Values of One-Track α_C and α_H for Balanced Mixed Designs (α = .05)

subjects per group	α_C	α_C as a function of α	α_H
2	.0129	α / 3.89	.0506
3	.0131	α / 3.82	.0511
4	.0133	α / 3.76	.0518
5	.0134	α / 3.72	.0523
6	.0136	α / 3.68	.0526
7	.0136	α / 3.66	.0529
8	.0137	α / 3.64	.0532
9	.0138	α / 3.63	.0533
10	.0138	α / 3.61	.0535
100	.0141	α / 3.54	.0544
1,000	.0142	α / 3.53	.0546
1,000,000,000	.0142	α / 3.53	.0546

Note. The underlying simulations used a population covariance of zero (i.e., nonnegative covariance was assumed).

APPENDIX G:

R CODE FOR `ac2x2between.R`

```
# These simulations find the optimal comparisonwise alpha level (aComp) for the effect
# tests in a 2x2 between-subjects design. The following structured-testing approach is
# used: First test the interaction at alpha; if it is significant, then go down the
# "simple-effect track" and test each simple effect at level aComp; otherwise, go down
# the "main-effect track" and test each main effect at level aComp. This program
# simulates experiments and does a binary search to find the optimal two-sided aComp
# for the simple-effect and main-effect tests that controls the experimentwise Type I
# error rate at the desired alpha level.

# Author: Andrew V. Frane (3/2019 using R version 3.3.3)

rm(list=ls())              # clear variables from workspace
startTime = proc.time()[3] # start timing this program

################################################################################

# INPUT PARAMETERS

numSim    = 10^7 # number of simulations
a         = .05  # familywise alpha level
n11       = 2    # number of observations in cell 11
n12       = 2    # number of observations in cell 12
n21       = 2    # number of observations in cell 21
n22       = 2    # number of observations in cell 22
poolIntSS = TRUE # whether interaction dropped from model when on main-effect track,
                 # i.e., whether to pool interaction SS w/ within-cell SS & add 1 df
                 # for main-effect tests

################################################################################

# FIXED PARAMETERS
# (can be changed to explore scenarios of unequal population means and/or variances)

mu11    = 0 # population mean for cell 11
mu12    = 0 # population mean for cell 12
mu21    = 0 # population mean for cell 21
mu22    = 0 # population mean for cell 22

sigma11 = 1 # population standard deviation for cell 11
sigma12 = 1 # population standard deviation for cell 12
sigma21 = 1 # population standard deviation for cell 21
sigma22 = 1 # population standard deviation for cell 22

################################################################################

# SIMULATIONS

# degrees of freedom
degf11 = n11 - 1
degf12 = n12 - 1
degf21 = n21 - 1
degf22 = n22 - 1
```

```
degf11v12 = degf11 + degf12
degf11v21 = degf11 + degf21
degf22v12 = degf22 + degf12
degf22v21 = degf22 + degf21
degf      = degf11 + degf12 + degf21 + degf22
degfx     = degf   + poolIntSS # for main-effect tests

# randomly generate sample means from normal distribution
means11 = rnorm(numSim, mu11, sigma11/sqrt(n11))
means12 = rnorm(numSim, mu12, sigma12/sqrt(n12))
means21 = rnorm(numSim, mu21, sigma21/sqrt(n21))
means22 = rnorm(numSim, mu22, sigma22/sqrt(n22))

# contrasts
emc1 = (n11+n12) / (n11*n12)
emc2 = (n21+n22) / (n21*n22)
emr1 = (n11+n21) / (n11*n21)
emr2 = (n12+n22) / (n12*n22)
cee  = emc2      / (emc1+emc2)
ree  = emr2      / (emr1+emr2)

meanMainC = cee * (means11-means12) + (1-cee) * (means21-means22) # 'cee' for column
meanMainR = ree * (means11-means21) + (1-ree) * (means12-means22) # 'ree' for row

meanInter = means11 - means12 - means21 + means22 # interaction
mean11v21 = means11 - means21
mean22v12 = means22 - means12
mean11v12 = means11 - means12
mean22v21 = means22 - means21

# sums of squares
ss11   = rchisq(numSim, degf11)
ss12   = rchisq(numSim, degf12)
ss21   = rchisq(numSim, degf21)
ss22   = rchisq(numSim, degf22)
ssPool = ss11 + ss12 + ss21 + ss22

# standard deviations
sd11v12 = sqrt((ss11+ss12) / degf11v12)
sd11v21 = sqrt((ss11+ss21) / degf11v21)
sd22v12 = sqrt((ss22+ss12) / degf22v12)
sd22v21 = sqrt((ss22+ss21) / degf22v21)
sdPool  = sqrt(ssPool      / degf)

# coefficients used to adjust numerator when computing t-statistics
adjInter = 1/sqrt(1/n11 + 1/n12 + 1/n21 + 1/n22)

adjMainC = 1/sqrt( cee^2 * (1/n11 + 1/n12) + (1-cee)^2 * (1/n21 + 1/n22) )
adjMainR = 1/sqrt( ree^2 * (1/n11 + 1/n21) + (1-ree)^2 * (1/n12 + 1/n22) )

adj11v12 = 1/sqrt(1/n11 + 1/n12)
adj11v21 = 1/sqrt(1/n11 + 1/n21)
adj22v12 = 1/sqrt(1/n22 + 1/n12)
adj22v21 = 1/sqrt(1/n22 + 1/n21)

# t-tests
tInter = adjInter * meanInter / sdPool

if (poolIntSS) {
  sdPoolInter = sqrt(((adjInter*meanInter)^2 + ssPool) / degfx )
```

```
  tMainC = adjMainC * meanMainC / sdPoolInter
  tMainR = adjMainR * meanMainR / sdPoolInter

} else {
  tMainC = adjMainC * meanMainC / sdPool
  tMainR = adjMainR * meanMainR / sdPool
}

t11v12 = adj11v12 * mean11v12 / sd11v12
t11v21 = adj11v21 * mean11v21 / sd11v21
t22v12 = adj22v12 * mean22v12 / sd22v12
t22v21 = adj22v21 * mean22v21 / sd22v21

# absolute-values of t-statistics
absTInter = abs(tInter)
absTMainC = abs(tMainC)
absTMainR = abs(tMainR)
absT11v12 = abs(t11v12)
absT11v21 = abs(t11v21)
absT22v12 = abs(t22v12)
absT22v21 = abs(t22v21)

################################################################################

# BINARY SEARCH FOR OPTIMAL COMPARISONWISE ALPHA LEVEL (aComp) FOR EFFECT TESTS

cpa      = -qt(a/2, degf)  # critical point of abs(t) for testing at level a
sigInter = absTInter > cpa # logical flag indicating whether interaction significant

aComp              = a/3             # initialize comparisonwise alpha level
targetNumSimsError = round(a*numSim) # target number of sims producing >=1 error

stepSize            = aComp/2 # initialize step-size to nudge candidate aComp up & dwn
aCompFound          = FALSE   # initialize flag indicating if optimal cpaComp found
latestStepDirection = 0       # initialize latest step direction (0=down, 1=up)

# "adjust" abs(t) stats to zero if given track not reached and/or given null is false
adjabsTMainC = absTMainC * !sigInter * (mu11+mu21 == mu12+mu22)
adjabsTMainR = absTMainR * !sigInter * (mu11+mu12 == mu21+mu22)

adjabsT11v12 = absT11v12 *  sigInter * (mu11 == mu12)
adjabsT11v21 = absT11v21 *  sigInter * (mu11 == mu21)
adjabsT22v12 = absT22v12 *  sigInter * (mu22 == mu12)
adjabsT22v21 = absT22v21 *  sigInter * (mu22 == mu21)

maxtMain = pmax(adjabsTMainC, adjabsTMainR) # max adjusted abs(t) stats for main FX

while (aCompFound==FALSE) {

  # comparisonwise critical values
  cpaCompdegfx     = -qt(aComp/2, degfx)
  cpaCompdegf11v12 = -qt(aComp/2, degf11v12)
  cpaCompdegf11v21 = -qt(aComp/2, degf11v21)
  cpaCompdegf22v12 = -qt(aComp/2, degf22v12)
  cpaCompdegf22v21 = -qt(aComp/2, degf22v21)

  # total number of simulations producing at least one Type I error
  numSimsError = sum(  (maxtMain     > cpaCompdegfx    )
                     | (adjabsT11v12 > cpaCompdegf11v12)
                     | (adjabsT11v21 > cpaCompdegf11v21)
                     | (adjabsT22v12 > cpaCompdegf22v12)
                     | (adjabsT22v21 > cpaCompdegf22v21) )
```

```
  # nudge aComp up or down as necessary
  if (numSimsError > targetNumSimsError) {          # ERROR RATE IS TOO HIGH
    stepSize = stepSize / (latestStepDirection + 1) # if prev step up, halve step size
    aComp    = max(0, aComp - stepSize)             # nudge down aComp
    latestStepDirection = 0                         # set latest step direction to dwn

  } else if (numSimsError < targetNumSimsError) {   # ERROR RATE IS TOO LOW
    stepSize = stepSize / (2 - latestStepDirection) # if prev step dwn, halve stepsize
    aComp    = aComp + stepSize                     # nudge up aComp
    latestStepDirection = 1                         # set latest step direction to up

  } else {                                          # ERROR RATE IS JUST RIGHT
    aCompFound = TRUE                               # optimal aComp found
  }
}

totalTime = (proc.time()[3] - startTime) / 60 # total time elapsed in minutes

################################################################################

# REPORT PARAMETERS AND RESULTS

totalTime

numSim                # number of simulations
a                     # experimentwise alpha level
c(n11, n12, n21, n22) # number of observations in each cell
poolIntSS             # whether interaction dropped from model on main-effect track
aComp                 # optimal comparisonwise alpha level
a/aComp               # ratio of experimentwise to comparisonwise alpha level
```

APPENDIX H:

R CODE FOR `ac2x2within.R`

```
# These simulations find the optimal comparisonwise alpha level (aComp) for effect
# tests in a 2x2 within-subjects design. The following structured testing approach is
# used: First test the interaction at alpha; if it is significant, then go down the
# "simple-effect track" and test each simple effect at aComp; otherwise, go down the
# "main-effect track" and test each main effect at aComp. This program simulates
# experiments and does a binary search to find the optimal two-sided aComp for the
# simple-effect and main-effect tests that controls the experimentwise Type I error
# rate at the desired alpha level.

# Author: Andrew V. Frane (3/2019 using R version 3.3.3)

rm(list=ls())              # clear variables from workspace
startTime = proc.time()[3] # start timing this program

# load 'MASS' package (for mvrnorm function)
if(!require(MASS)) {       # try to load package; and if it isn't found...
  install.packages("MASS") # ...install it...
  require(MASS)            # ...and load it
}

################################################################################

# INPUT PARAMETERS

numSim = 10^7 # number of simulations
a      = .05  # experimentwise alpha level
nSubj  = 2    # number of subjects

################################################################################

# FIXED PARAMETERS
# (can be changed to explore scenarios of unequal population means and/or variances)

mu11 = 0 # standardized population mean for cell 11
mu12 = 0 # standardized population mean for cell 12
mu21 = 0 # standardized population mean for cell 21
mu22 = 0 # standardized population mean for cell 22

popCov = matrix(c(1, 0, 0, 0,
                  0, 1, 0, 0,
                  0, 0, 1, 0,
                  0, 0, 0, 1), ncol=4) # population covariance matrix

################################################################################

# SIMULATIONS

mu   = c(mu11, mu12, mu21, mu22) # vector of population means
degf = nSubj - 1                 # degrees of freedom

# generate sample means and covariance matrices:
# in each "means" matrix of sample means, each row is a simulation and
#                                         each column is a cell: 11, 12, 21, 22.
```

```
# in each "covs" array, each slice in 3rd dimension is a 4x4 covariance matrix for
# given simulation.

# if nSubj>4, generate covariance matrices directly from the Wishart distribution
# (see Anderson 1958, An Introduction to Multivariate Statistical Analysis, p. 159,
# Theorem 7.2.2) and generate means directly from multivariate normal distribution.

# if nSubj<5 , generate random observations from multivariate normal distribution:
# observations start as a "temp1" matrix in which rows are subjects and columns are
# cells, then are reshaped into a "temp2" array in which each slice is a 4-by-n matrix
# for given simulation, then are reshaped into a "y" array in which each slice is an
# n-by-4 matrix for a given simulation. then the means matrix and covs array can be
# computed from the observations.

if (nSubj > 4) {
  means = mvrnorm( numSim, mu  , popCov/nSubj)  # means matrix
  covs  = rWishart(numSim, degf, popCov) / degf # covs matrix

} else {
  yTemp1 = mvrnorm(numSim*nSubj, mu, popCov)               # temp1 matrix
  yTemp2 = array(t(yTemp1), c(4, nSubj, numSim))           # temp2 array
  y      = array(apply(yTemp2, 3, t), c(nSubj, 4, numSim)) # y array

  means  = apply(y, c(3, 2), mean)                         # means matrix
  covs   = array(apply(y, 3, cov), c(4, 4, numSim))        # covs array
}

# define contrasts
contrastInter = c(1, -1, -1,  1)
contrastMain1 = c(1,  1, -1, -1)
contrastMain2 = c(1, -1,  1, -1)
contrast11v12 = c(1, -1,  0,  0)
contrast11v21 = c(1,  0, -1,  0)
contrast22v12 = c(0, -1,  0,  1)
contrast22v21 = c(0,  0, -1,  1)

# compute means for contrasts
meanInter = means %*% contrastInter
meanMain1 = means %*% contrastMain1
meanMain2 = means %*% contrastMain2
mean11v12 = means %*% contrast11v12
mean11v21 = means %*% contrast11v21
mean22v12 = means %*% contrast22v12
mean22v21 = means %*% contrast22v21

# define functions for computing quadratic forms from given covariance matrix
quadFormInter = function(covX) t(contrastInter) %*% covX %*% contrastInter
quadFormMain1 = function(covX) t(contrastMain1) %*% covX %*% contrastMain1
quadFormMain2 = function(covX) t(contrastMain2) %*% covX %*% contrastMain2
quadForm11v12 = function(covX) t(contrast11v12) %*% covX %*% contrast11v12
quadForm11v21 = function(covX) t(contrast11v21) %*% covX %*% contrast11v21
quadForm22v12 = function(covX) t(contrast22v12) %*% covX %*% contrast22v12
quadForm22v21 = function(covX) t(contrast22v21) %*% covX %*% contrast22v21

# compute standard errors for contrasts
semInter = sqrt(apply(covs, 3, quadFormInter))
semMain1 = sqrt(apply(covs, 3, quadFormMain1))
semMain2 = sqrt(apply(covs, 3, quadFormMain2))
sem11v12 = sqrt(apply(covs, 3, quadForm11v12))
sem11v21 = sqrt(apply(covs, 3, quadForm11v21))
sem22v12 = sqrt(apply(covs, 3, quadForm22v12))
```

```
sem22v21 = sqrt(apply(covs, 3, quadForm22v21))

# compute t-statistics
tInter = sqrt(nSubj) * meanInter / semInter
tMain1 = sqrt(nSubj) * meanMain1 / semMain1
tMain2 = sqrt(nSubj) * meanMain2 / semMain2
t11v12 = sqrt(nSubj) * mean11v12 / sem11v12
t11v21 = sqrt(nSubj) * mean11v21 / sem11v21
t22v12 = sqrt(nSubj) * mean22v12 / sem22v12
t22v21 = sqrt(nSubj) * mean22v21 / sem22v21

# absolute value of t-statistics
absTInter = abs(tInter)
absTMain1 = abs(tMain1)
absTMain2 = abs(tMain2)
absT11v12 = abs(t11v12)
absT11v21 = abs(t11v21)
absT22v12 = abs(t22v12)
absT22v21 = abs(t22v21)

#############################################################################################

# BINARY SEARCH FOR OPTIMAL COMPARISONWISE ALPHA LEVEL (aComp) FOR EFFECT TESTS

if ( (mu11+mu12 != mu21+mu22) & (mu11 != mu12) & (mu11 != mu21) &
     (mu11+mu21 != mu12+mu22) & (mu22 != mu12) & (mu22 != mu21) ) {

  cpaComp = 0 # consider optimal comparisonwise crit value 0 if all effect nulls false
  aComp   = 1 # consider optimal comparisonwise alphalevel 1 if all effect nulls false

} else {
  cpa      = -qt(a/2, degf)  # critical point of abs(t) for testing at level a
  sigInter = absTInter > cpa # logical flags indicating if interaction was significant

  # find max abs(t) in each sim; set to 0 if null false and/or given track not taken
  maxTMain  = pmax(absTMain1*(mu11+mu12==mu21+mu22),
                   absTMain2*(mu11+mu21==mu12+mu22)) * !sigInter

  maxTSimp  = pmax(absT11v12*(mu11==mu12), absT11v21*(mu11==mu21),
                   absT22v12*(mu22==mu12),
                   absT22v21*(mu22==mu21)) *  sigInter

  maxTGrand = pmax(maxTMain,maxTSimp)

  targetNumSimsError = round(a*numSim) # target number of sims with >=1 error
  cpaComp            = -qt(a/6, degf)  # initialize comparisonwise critical value
  stepSize           = cpaComp/2       # initialize step size to nudge value up & down

  cpaCompFound        = FALSE # initialize flag indicating if optimal crit value found
  latestStepDirection = 1     # initialize latest step direction (0=down, 1=up)

  while (cpaCompFound==FALSE) { # stay in while-loop until optimal crit value is found

    numSimsError = sum(maxTGrand > cpaComp) # number of sims with >=1 Type I error

    # nudge cpaComp up or down as necessary
    if (numSimsError < targetNumSimsError) {        # ERROR RATE IS TOO LOW
      stepSize = stepSize / (latestStepDirection+1) # if prev step up, halve step size
      cpaComp = cpaComp - stepSize                  # nudge down critical value
      latestStepDirection = 0                       # set latest step direction to dwn
```

```
    } else if (numSimsError > targetNumSimsError) { # ERROR RATE IS TOO HIGH
      stepSize = stepSize / (2-latestStepDirection) # if prev step dwn, halve stepsize
      cpaComp = cpaComp + stepSize                  # nudge up critical value
      latestStepDirection = 1                       # set latest step direction to up

    } else {                                        # ERROR RATE IS JUST RIGHT
      cpaCompFound = TRUE                           # optimal critical value found
    }
  }
}

aComp     = 2*pt(-cpaComp, degf)              # optimal comparisonwise alpha level
totalTime = (proc.time()[3] - startTime) / 60 # total time elapsed in minutes

################################################################################

# REPORT PARAMETERS AND RESULTS

totalTime

numSim # number of simulations
a      # experimentwise alpha level
nSubj  # number of subjects
aComp  # optimal comparisonwise alpha level
```

APPENDIX I:

R CODE FOR `ac2x2mixed.R`

```
# These simulations find the optimal comparisonwise alpha level (aComp) for effect
# tests in a 2x2 mixed design. The following structured testing approach is used:
# First test the interaction at alpha; if it is significant, then go down the "simple
# effect track" and test each simple effect at aComp; otherwise, go down the "main
# effect track" and test each main effect at aComp. This program simulates experiments
# and does a binary search to find the optimal two-sided aComp for the simple-effect
# and main-effect tests that controls the experimentwise Type I error rate at the
# desired alpha level.

# Author: Andrew V. Frane (3/2019 using R version 3.3.3)

rm(list=ls())              # clear variables from workspace
startTime = proc.time()[3] # start timing this program

# load 'MASS' package (for mvrnorm function)
if(!require(MASS)) {       # try to load package; and if it isn't found...
  install.packages("MASS") # ...install it...
  require(MASS)            # ...and load it
}

##############################################################################################

# INPUT PARAMETERS

numSim    = 10^7 # number of simulations
a         = .05  # experimentwise alpha level
rho       = 0    # population covariance for within-subjects factor
                 # (can be 0 if nonnegativity assumed)

nGroup1   = 10   # number of subjects in group 1
nGroup2   = 10   # number of subjects in group 2

poolIntSS = TRUE # whether interaction is dropped from model on main-effect track,
                 # i.e., whether to pool the interaction SS with within-cell SS & add
                 # 1 df for the within-subjects main-effect test

##############################################################################################

# FIXED PARAMETERS
# (can be changed to explore scenarios of unequal population means and/or covariance
# matrices)

mu11 = 1 # population mean for level 1 of within-subjects factor in group 1
mu12 = 1 # population mean for level 2 of within-subjects factor in group 1
mu21 = 1 # population mean for level 1 of within-subjects factor in group 2
mu22 = 1 # population mean for level 2 of within-subjects factor in group 2

popCov1112 = matrix(c( 1 ,rho,
                       rho, 1), ncol=2) # population covariance matrix for group 1

popCov2122 = matrix(c( 1, rho,
                       rho, 1), ncol=2) # population covariance matrix for group 2
```

```
#################################################################################

# SIMULATIONS

# degrees of freedom
degf1 = nGroup1 - 1
degf2 = nGroup2 - 1
degf  = degf1   + degf2
degfx = degf    + poolIntSS

# generate sample means and covariance matrices:
# in each "means" matrix of sample means, each row is a simulation and
#                                         each column is a cell: 11, 12, 21, 22.

# in each "covs" array, each slice in the 3rd dimension is a 2x2 covariance matrix for
# a given simulation.

# if n>2 for given group, generate covariance matrices directly from the Wishart
# distribution (see Anderson 1958, An Introduction to Multivariate Statistical
# Analysis, p. 159, Theorem 7.2.2) and generate means directly from multivariate
# normal distribution.

# if n<3 for given group, generate random observations from multivariate normal
# distribution: the observations start as a "temp1" matrix in which rows are subjects
# and columns are cells, then are reshaped into a "temp2" array in which each slice is
# a 2-by-n matrix for given simulation, then are reshaped into a "y" array in which
# each slice is an n-by-2 matrix for a given simulation. then the means matrix and
# covs array can be computed from the observations.

if (nGroup1 > 2) {

  means1112 = mvrnorm(numSim, c(mu11, mu12), popCov1112 / nGroup1)    # means mtrx grp1
  covs1112  = rWishart(numSim, degf1, popCov1112) / degf1             # covs array grp1

} else {

  y1112Temp1 = mvrnorm(numSim*nGroup1, c(mu11, mu12), popCov1112)     # temp1 mtrx grp1
  y1112Temp2 = array(t(y1112Temp1), c(2, nGroup1, numSim))            # temp2 arry grp1
  y1112      = array(apply(y1112Temp2, 3, t), c(nGroup1, 2, numSim)) # y array group 1
  means1112  = apply(y1112, c(3, 2), mean)                           # means mtrx grp1
  covs1112   = array(apply(y1112, 3, cov), c(2, 2, numSim))          # covs array grp1
}

if (nGroup2 > 2) {
  means2122 = mvrnorm(numSim, c(mu21, mu22), popCov2122 / nGroup2)    # means mtrx grp2
  covs2122  = rWishart(numSim, degf2, popCov2122) / degf2             # covs array grp2

} else {
  y2122Temp1 = mvrnorm(numSim*nGroup2, c(mu21, mu22), popCov2122)     # temp1 mtrx grp2
  y2122Temp2 = array(t(y2122Temp1), c(2, nGroup2, numSim))            # temp2 arry grp2
  y2122      = array(apply(y2122Temp2, 3, t), c(nGroup2, 2, numSim)) # y array group 2
  means2122  = apply(y2122, c(3, 2), mean)                           # means mtrx grp2
  covs2122   = array(apply(y2122, 3, cov), c(2, 2, numSim))          # covs array grp2
}

# sample pooled within-group 2x2 covariance matrix for repeated measures
covsPool = (degf1*covs1112 + degf2*covs2122) / degf

# define contrasts for mean comparisons
nAvg = (nGroup1 + nGroup2) / 2
f    =  nGroup1 / nAvg
g    =  nGroup2 / nAvg
```

```
contrastInter = c( 1, -1, -1,  1) # interaction
contrastMainB = c( 1,  1, -1, -1) # between-groups main effect
contrastMainW = c( f, -f,  g, -g) # within-groups main effect weighted by group sizes
contrast11v12 = c( 1, -1)         # simple effect within group 1
contrast22v21 = c(-1,  1)         # simple effect within group 2
contrast11v21 = c( 1,  0, -1,  0) # simple effect between groups for first measure
contrast22v12 = c( 0, -1,  0,  1) # simple effect between groups for second measure

# compute sample mean differences
means     = cbind(means1112, means2122)
meanInter = means     %*% contrastInter
meanMainB = means     %*% contrastMainB
meanMainW = means     %*% contrastMainW
mean11v21 = means     %*% contrast11v21
mean22v12 = means     %*% contrast22v12
mean11v12 = means1112 %*% contrast11v12
mean22v21 = means2122 %*% contrast22v21

adjOne = sqrt(nGroup1+nGroup2) / 2
adjTwo = sqrt(nGroup1*nGroup2  / (nGroup1+nGroup2))
adjGr1 = sqrt(nGroup1)
adjGr2 = sqrt(nGroup2)

# contrasts for sample covariance matrices
contrastDif = c(1, -1)
contrastSum = c(1,  1)

# define quadratic forms used to compute variances
quadFormSum   = function(covX) t(contrastSum)   %*% covX %*% contrastSum
quadFormDif   = function(covX) t(contrastDif)   %*% covX %*% contrastDif
quadForm11v12 = function(covX) t(contrast11v12) %*% covX %*% contrast11v12
quadForm22v21 = function(covX) t(contrast22v21) %*% covX %*% contrast22v21
quadFormVar1  = function(covX) covX[1, 1]
quadFormVar2  = function(covX) covX[2, 2]

# compute variances
varDif   = apply(covsPool, 3, quadFormDif)
varMainB = apply(covsPool, 3, quadFormSum)
var11v21 = apply(covsPool, 3, quadFormVar1)
var22v12 = apply(covsPool, 3, quadFormVar2)
var11v12 = apply(covs1112, 3, quadForm11v12)
var22v21 = apply(covs2122, 3, quadForm22v21)

varMainW = (varDif*degf + ((adjOne*meanInter)^2)*poolIntSS) / degfx

# compute t-statistics
tInter = adjTwo * meanInter / sqrt(varDif  ) # 2-sample t-test of interaction
tMainB = adjTwo * meanMainB / sqrt(varMainB) # 2-sample t-test btwn-subjs main effect
tMainW = adjOne * meanMainW / sqrt(varMainW) # 1-sample t-test wthn-subjs main effect
t11v12 = adjGr1 * mean11v12 / sqrt(var11v12) # 1-sample t-test within group 1
t22v21 = adjGr2 * mean22v21 / sqrt(var22v21) # 1-sample t-test within group 2
t11v21 = adjTwo * mean11v21 / sqrt(var11v21) # 2-sample t-test between groups
                                             # at level 1 of within-group factor
t22v12 = adjTwo * mean22v12 / sqrt(var22v12) # 2-sample t-test between groups
                                             # at level 2 of within-group factor

# absolute value of t-statistics
absTInter = abs(tInter)
absTMainB = abs(tMainB)
absTMainW = abs(tMainW)
absT11v12 = abs(t11v12)
```

```
absT11v21 = abs(t11v21)
absT22v12 = abs(t22v12)
absT22v21 = abs(t22v21)

################################################################################

# BINARY SEARCH FOR OPTIMAL COMPARISONWISE ALPHA LEVEL (aComp) FOR EFFECT TESTS

if (mu11+mu12 != mu22+mu21 & mu11 != mu12  &  mu11 != mu21  &
    mu11+mu21 != mu22+mu12 & mu22 != mu12  &  mu22 != mu21) {

  cpaComp = 0 # consider optimal comparisonwise crit value 0 if all effect nulls false
  aComp   = 1 # consider optimal comparisonwise alphalevel 1 if all effect nulls false

} else {
  aComp              = a/3             # starting value for aComp
  targetNumSimsError = round(a*numSim) # target number of sims with >=1 error

  stepSize            = aComp/2 # initialize stepsize to nudge candidate aComp up&down
  aCompFound          = FALSE   # initialize flag indicating if optimal cpaComp found
  latestStepDirection = 0       # initialize latest step direction (0=down, 1=up)

  cpaInter = -qt(a/2, degf)       # critical point of abs(t) for interaction test
  sigInter = absTInter > cpaInter # logical flags indicating if interaction was signif

  # "adjust" t-statistics to zero when given track not reached and/or when null
  # hypothesis is false
  adjAbsTMainB = absTMainB * !sigInter * (mu11+mu12==mu21+mu21)
  adjAbsTMainW = absTMainW * !sigInter * (mu11+mu21==mu22+mu12)
  adjAbsT11v12 = absT11v12 *  sigInter * (mu11==mu12)
  adjAbsT11v21 = absT11v21 *  sigInter * (mu11==mu21)
  adjAbsT22v12 = absT22v12 *  sigInter * (mu22==mu12)
  adjAbsT22v21 = absT22v21 *  sigInter * (mu22==mu21)

  # maximum abs t-stat out of all between-group effect tests that have true nulls and
  # were on the track that was reached
  maxTdegf = pmax(adjAbsTMainB, adjAbsT11v21, adjAbsT22v12)

  while (aCompFound==FALSE) {

    # critical points
    cpaCompdegf  = -qt(aComp/2, degf)  # for between-group main effect
    cpaCompdegfx = -qt(aComp/2, degfx) # for within-group main effect
    cpaCompdegf1 = -qt(aComp/2, degf1) # for simple effect within group 1
    cpaCompdegf2 = -qt(aComp/2, degf2) # for simple effect within group 2

    # total number of simulations producing at least one Type I error
    numSimsError = sum((maxTdegf     > cpaCompdegf ) | (adjAbsTMainW > cpaCompdegfx) |
                       (adjAbsT11v12 > cpaCompdegf1) | (adjAbsT22v21 > cpaCompdegf2))

    # nudge comparisonwise alpha level up or down as necessary
    if (numSimsError > targetNumSimsError) {        # ERROR RATE TOO HIGH
      stepSize = stepSize / (latestStepDirection+1) # if prev step up, halve step size
      aComp               = aComp - stepSize        # nudge down aComp
      latestStepDirection = 0                       # set latest step direction to dwn

    } else if (numSimsError < targetNumSimsError) { # ERROR RATE TOO LOW
      stepSize = stepSize / (2-latestStepDirection) # if prev step dwn, halve stepsize

      aComp               = aComp + stepSize        # nudge up aComp
      latestStepDirection = 1                       # set latest step direction to up
```

```
    } else {                                          # ERROR RATE IS JUST RIGHT
      aCompFound = TRUE                               # optimal aComp found
    }
  }
}

totalTime = (proc.time()[3] - startTime) / 60  # total time elapsed in minutes

################################################################################

# REPORT RESULTS

totalTime

numSim    # number of simulations
a         # experimentwise alpha level
nGroup1   # number of subjects in group 1
nGroup2   # number of subjects in group 2
poolIntSS # whether interaction dropped from model on main-effect track
aComp     # optimal comparisonwise alpha level
```

APPENDIX J:

R CODE FOR `sim2x2between.R`

```
# These simulations compare any-test power and experimentwise error rates (EWERs) of
# different methods in a 2x2 between-subjects design. The "two-track" methods test
# only the main effects when the interaction is nonsignificant, and test only the
# simple (pairwise) effects when the interaction is significant. The "one-track"
# methods test only the simple-effect tests (without considering the interaction or
# main effects).

# Author: Andrew V. Frane (4/2019 using R version 3.3.3)

rm(list=ls())              # clear variables from workspace
startTime = proc.time()[3] # start timing this program

################################################################################

# INPUT PARAMETERS

numSim    = 10^6  # number of simulations
a         = .05   # familywise alpha level

aComp1    = .0141 # comparisonwise alpha for one-track AC method;
                  # obtain using 'ac2x2betweensimp.R'

aComp2    = .0144 # comparisonwise alpha for two-track AC method;
                  # obtain using 'ac2x2between.R'

aHom      = .0545 # nominal familywise alpha for one-track Hommel method;
                  # obtain using 'ac2x2betweensimp.R'

n11       = 50    # number of observations in cell 11
n12       = 50    # number of observations in cell 12
n21       = 50    # number of observations in cell 21
n22       = 50    # number of observations in cell 22

mu11      = 0     # population mean for cell 11
mu12      = .3    # population mean for cell 12
mu21      = .3    # population mean for cell 21
mu22      = .7    # population mean for cell 22

sigma11   = 1     # population standard deviation for cell 11
sigma12   = 1     # population standard deviation for cell 12
sigma21   = 1     # population standard deviation for cell 21
sigma22   = 1     # population standard deviation for cell 22

poolIntSS = TRUE  # whether interatction dropped from model when on main-effect track

################################################################################

# SIMULATIONS

# degrees of freedom
degf11 = n11 - 1
degf12 = n12 - 1
```

```
degf21 = n21 - 1
degf22 = n22 - 1

degf11v12 = degf11 + degf12
degf11v21 = degf11 + degf21
degf22v12 = degf22 + degf12
degf22v21 = degf22 + degf21
degf      = degf11 + degf12 + degf21 + degf22
degfx     = degf   + poolIntSS # for main-effect tests

# randomly generate sample means from normal distribution
means11 = rnorm(numSim, mu11, sigma11/sqrt(n11))
means12 = rnorm(numSim, mu12, sigma12/sqrt(n12))
means21 = rnorm(numSim, mu21, sigma21/sqrt(n21))
means22 = rnorm(numSim, mu22, sigma22/sqrt(n22))

# contrasts
emc1 = (n11+n12) / (n11*n12)
emc2 = (n21+n22) / (n21*n22)
emr1 = (n11+n21) / (n11*n21)
emr2 = (n12+n22) / (n12*n22)
cee  = emc2      / (emc1+emc2)
ree  = emr2      / (emr1+emr2)

meanMainC = cee * (means11-means12) + (1-cee) * (means21-means22) # 'cee' for column
meanMainR = ree * (means11-means21) + (1-ree) * (means12-means22) # 'ree' for row

meanInter = means11 - means12 - means21 + means22
mean11v12 = means11 - means12
mean11v21 = means11 - means21
mean22v12 = means22 - means12
mean22v21 = means22 - means21

# sums of squares
ss11   = rchisq(numSim, degf11)
ss12   = rchisq(numSim, degf12)
ss21   = rchisq(numSim, degf21)
ss22   = rchisq(numSim, degf22)
ssPool = ss11 + ss12 + ss21 + ss22

# standard deviations
sd11v12 = sqrt((ss11+ss12) / degf11v12)
sd11v21 = sqrt((ss11+ss21) / degf11v21)
sd22v12 = sqrt((ss22+ss12) / degf22v12)
sd22v21 = sqrt((ss22+ss21) / degf22v21)
sdPool  = sqrt(ssPool      / degf)

# coefficients used to adjust numerator when computing t-statistics
adjInter = 1/sqrt(1/n11 + 1/n12 + 1/n21 + 1/n22)

adjMainC = 1/sqrt( cee^2 * (1/n11 + 1/n12) + (1-cee)^2 * (1/n21 + 1/n22) )
adjMainR = 1/sqrt( ree^2 * (1/n11 + 1/n21) + (1-ree)^2 * (1/n12 + 1/n22) )

adj11v12 = 1/sqrt(1/n11 + 1/n12)
adj11v21 = 1/sqrt(1/n11 + 1/n21)
adj22v12 = 1/sqrt(1/n22 + 1/n12)
adj22v21 = 1/sqrt(1/n22 + 1/n21)

# t-statistics
tInter = adjInter * meanInter / sdPool
```

```
if (poolIntSS) {
  sdPoolInter = sqrt( ((adjInter*meanInter)^2 + ssPool) / degfx )

  tMainC = adjMainC * meanMainC / sdPoolInter
  tMainR = adjMainR * meanMainR / sdPoolInter

} else {
  tMainC = adjMainC * meanMainC / sdPool
  tMainR = adjMainR * meanMainR / sdPool
}

t11v12 = adj11v12 * mean11v12 / sd11v12
t11v21 = adj11v21 * mean11v21 / sd11v21
t22v12 = adj22v12 * mean22v12 / sd22v12
t22v21 = adj22v21 * mean22v21 / sd22v21

# absolute-value t-statistics
absTInter = abs(tInter)
absTMainC = abs(tMainC)
absTMainR = abs(tMainR)
absT11v12 = abs(t11v12)
absT11v21 = abs(t11v21)
absT22v12 = abs(t22v12)
absT22v21 = abs(t22v21)

# test the interaction
cpadegf  = -qt(a/2, degf)      # unadjusted crit point for 2-sided test with df=degf
sigInter = absTInter > cpadegf # flag significant interaction tests

# p-values for simple-effect t-statistics
p11v12 = 2*pt(-absT11v12, degf11v12)
p11v21 = 2*pt(-absT11v21, degf11v21)
p22v12 = 2*pt(-absT22v12, degf22v12)
p22v21 = 2*pt(-absT22v21, degf22v21)

# matrix of p-values (each row is for a given simulation)
pMat = cbind(p11v12, p11v21, p22v12, p22v21)

# null hypothesis statuses
nullMainC = (mu11+mu21 == mu12+mu22)
nullMainR = (mu11+mu12 == mu21+mu22)
null11v12 = (mu11 == mu12)
null11v21 = (mu11 == mu21)
null22v12 = (mu22 == mu12)
null22v21 = (mu22 == mu21)

# number of simple false nulls
numSimpleFalseNull = sum(!null11v12, !null11v21, !null22v12, !null22v21)

# are all nulls true?
nullAll = (numSimpleFalseNull == 0)

# flag significant tests for one-track AC method
sig11v12AC1 = (p11v12 < aComp1)
sig11v21AC1 = (p11v21 < aComp1)
sig22v12AC1 = (p22v12 < aComp1)
sig22v21AC1 = (p22v21 < aComp1)

# flag significant tests for two-track AC method
cpaComp2degfx = -qt(aComp2/2, degfx) # crit point for main effects at level aComp2
```

```
sigMainCAC2 = (absTMainC > cpaComp2degfx) & !sigInter
sigMainRAC2 = (absTMainR > cpaComp2degfx) & !sigInter


sig11v12AC2 = (p11v12 < aComp2) & sigInter
sig11v21AC2 = (p11v21 < aComp2) & sigInter
sig22v12AC2 = (p22v12 < aComp2) & sigInter
sig22v21AC2 = (p22v21 < aComp2) & sigInter

# flag significant tests for two-track method with no adjustment
cpadegfx = -qt(a/2, degfx) # crit point for main effects at unadjusted alpha level

sigMainCNoAdj = (absTMainC > cpadegfx) & !sigInter
sigMainRNoAdj = (absTMainR > cpadegfx) & !sigInter

sig11v12NoAdj = (p11v12 < a) & sigInter
sig11v21NoAdj = (p11v21 < a) & sigInter
sig22v12NoAdj = (p22v12 < a) & sigInter
sig22v21NoAdj = (p22v21 < a) & sigInter

# flag significant tests for 2-track "given-track" Bonferroni
# (adjusts for tests on reached track only)
cpaBonfMain = -qt(a/4, degfx) # crit point for main effects at level alpha/2

sigMainCBonf = (absTMainC > cpaBonfMain) & !sigInter
sigMainRBonf = (absTMainR > cpaBonfMain) & !sigInter

sig11v12Bonf = (p11v12 < a/4) & sigInter
sig11v21Bonf = (p11v21 < a/4) & sigInter
sig22v12Bonf = (p22v12 < a/4) & sigInter
sig22v21Bonf = (p22v21 < a/4) & sigInter

# flag significant tests for one-track Hommel method
hommelAdjust = function(p) p.adjust(p, "hom") # function to do Hommel adjustments
pHom         = apply(pMat, 1, hommelAdjust)   # matrix of hommel-adjusted p-values
                                              # (each column is for a given sim)

sig11v12Hom = pHom[1,] < aHom
sig11v21Hom = pHom[2,] < aHom
sig22v12Hom = pHom[3,] < aHom
sig22v21Hom = pHom[4,] < aHom

# EWER for one-track AC method
anySigAC1TrueNull = (sig11v12AC1 & null11v12) | (sig11v21AC1 & null11v21) |
                    (sig22v12AC1 & null22v12) | (sig22v21AC1 & null22v21)

ewerAC1 = mean(anySigAC1TrueNull)

# EWER for two-track AC method
anySigAC2TrueNull = (sigMainCAC2 & nullMainC) | (sigMainRAC2 & nullMainR) |
                    (sig11v12AC2 & null11v12) | (sig11v21AC2 & null11v21) |
                    (sig22v12AC2 & null22v12) | (sig22v21AC2 & null22v21)

ewerAC2 = mean(anySigAC2TrueNull)

# EWER for two-track approach with no adjustment
anySigNoAdjTrueNull = (sigMainCNoAdj & nullMainC) | (sigMainRNoAdj & nullMainR) |
                      (sig11v12NoAdj & null11v12) | (sig11v21NoAdj & null11v21) |
                      (sig22v12NoAdj & null22v12) | (sig22v21NoAdj & null22v21)

ewerNoAdj = mean(anySigNoAdjTrueNull)
```

```
# EWER for two-track "given-track" Bonferroni method
# (adjusts for tests on reached track only)
anySigBonfTrueNullGiven = (sigMainCBonf & nullMainC) | (sigMainRBonf & nullMainR) |
                          (sig11v12Bonf & null11v12) | (sig11v21Bonf & null11v21) |
                          (sig22v12Bonf & null22v12) | (sig22v21Bonf & null22v21)

ewerBonfGiven = mean(anySigBonfTrueNullGiven)

# EWER for two-track "partial" Bonferroni method (only adjusts on simple-effect track)
anySigBonfTrueNullPart = (sigMainCNoAdj & nullMainC) | (sigMainRNoAdj & nullMainR) |
                         (sig11v12Bonf  & null11v12) | (sig11v21Bonf  & null11v21) |
                         (sig22v12Bonf  & null22v12) | (sig22v21Bonf  & null22v21)

ewerBonfPart = mean(anySigBonfTrueNullPart)

# "de facto" EWER for two-track AC method (considers main effects as Type I errors if
# null for either pairwise test of given factor is true)
anySigACTrueNullDF = anySigAC2TrueNull | (sigMainCAC2 & (null11v12 | null22v21)) |
                                         (sigMainRAC2 & (null11v21 | null22v12))

ewerACDF = mean(anySigACTrueNullDF)

# EWER for one-track Hommel method
anySigHomTrueNull = (sig11v12Hom & null11v12) | (sig11v21Hom & null11v21) |
                    (sig22v12Hom & null22v12) | (sig22v21Hom & null22v21)

ewerHom = mean(anySigHomTrueNull)

# compute power
if (nullAll) {
  avgPowerAC1 = NaN
  avgPowerHom = NaN
  anyPowerAC1 = NaN
  anyPowerAC2 = NaN
  anyPowerHom = NaN

} else {

  # per-test power for one-track AC method
  numSigAC1FalseNull = (sig11v12AC1 & !null11v12) + (sig11v21AC1 & !null11v21) +
                       (sig22v12AC1 & !null22v12) + (sig22v21AC1 & !null22v21)

  avgPowerAC1 = mean(numSigAC1FalseNull / numSimpleFalseNull)

  # any-test power for one-track AC method
  anySigAC1FalseNull = numSigAC1FalseNull > 0
  anyPowerAC1        = mean(anySigAC1FalseNull)

  # any-test power for two-track AC method
  anySigAC2FalseNull = (sigMainCAC2 & !nullMainC) | (sigMainRAC2 & !nullMainR) |
                       (sig11v12AC2 & !null11v12) | (sig11v21AC2 & !null11v21) |
                       (sig22v12AC2 & !null22v12) | (sig22v21AC2 & !null22v21)

  anyPowerAC2 = mean(anySigAC2FalseNull)

  # per-test power for one-track Hommel method
  numSigHomFalseNull = (sig11v12Hom & !null11v12) + (sig11v21Hom & !null11v21) +
                       (sig22v12Hom & !null22v12) + (sig22v21Hom & !null22v21)

  avgPowerHom = mean(numSigHomFalseNull / numSimpleFalseNull)
```

```
  # any-test power for one-track Hommel method
  anySigHomFalseNull = numSigHomFalseNull > 0
  anyPowerHom        = mean(anySigHomFalseNull)
}

totalTime = (proc.time()[3] - startTime) / 60 # total time elapsed in minutes

################################################################################

# REPORT PARAMETERS AND RESULTS

totalTime

numSim                                  # number of simulations
a                                       # experimentwise alpha level
aComp1                                  # comparisonwise alpha level for one-track AC
aComp2                                  # comparisonwise alpha level for two-track AC
aHom                                    # nominal familywise alpha for one-track Hommel
c(    n11,     n12,     n21,     n22) # number of observations in each cell
c(   mu11,    mu12,    mu21,    mu22) # population cell means
c(sigma11, sigma12, sigma21, sigma22) # population cell standard deviations
poolIntSS                               # whether interaction dropped from model when on
                                        # main-effect track

avgPowerAC1   # per-test power for the one-track aComp  method
avgPowerHom   # per-test power for the one-track Hommel method

anyPowerAC1   # any-test power for the one-track aComp  method
anyPowerAC2   # any-test power for the two-track aComp  method
anyPowerHom   # any-test power for the one-track Hommel method

ewerAC1       # experimentwise Type I error rate for the one-track aComp method
ewerAC2       # experimentwise Type I error rate for the two-track aComp method
ewerNoAdj     # experimentwise Type I error rate for the two-track no-adjustment
method
ewerBonfGiven # experimentwise Type I error rate for two-track given-track Bonferroni
ewerBonfPart  # experimentwise Type I error rate for two-track partial Bonferroni
ewerACDF      # "de facto" experimentwise Type I error rate for two-track aComp method
ewerHom       # experimentwise Type I error rate for one-track Hommel method
```

APPENDIX K:

R CODE FOR `sim2x2within.R`

```
# These simulations compare any-test power and experimentwise error rates (EWERs) of
# different methods in a 2x2 within subjects design. The "two-track" methods test only
# the main effects when the interaction is nonsignificant, and test only the simple
# (pairwise) effects when the interaction is significant. The "one-track" methods test
# only the simple-effect tests (without considering the interaction or main effects).

# Author: Andrew V. Frane (4/2019 using R version 3.3.3)

rm(list=ls())              # clear variables from workspace
startTime = proc.time()[3] # start timing this program

# load 'MASS' package (for mvrnorm function)
if(!require(MASS)) {       # try to load package; and if it isn't found...
  install.packages("MASS") # ...install it...
  require(MASS)            # ...and load it
}

##############################################################################

# INPUT PARAMETERS

numSim = 10^6  # number of simulations
a      = .05   # experimentwise alpha level

aComp1 = .0132 # comparisonwise alpha for one-track AC method;
               # obtain using 'ac2x2withinsimp.R'

aComp2 = .0174 # comparisonwise alpha for two-track AC method;
               # obtain using 'ac2x2within.R'

aHom   = .0515 # nominal familywise alpha level for one-track Hommel method;
               # obtain using 'ac2x2withinsimp.R'

nSubj  = 5     # number of subjects

# standardized population means
mu11 = 0
mu12 = 0
mu21 = 0
mu22 = 0

# population covariance
rho = 0

##############################################################################

# SIMULATIONS

mu   = c(mu11, mu12, mu21, mu22) # vector of population means
degf = nSubj - 1                 # degrees of freedom
```

```
popCov = matrix(c(  1 , rho, rho, rho,
                  rho,  1 , rho, rho,
                  rho, rho,  1 , rho,
                  rho, rho, rho,  1), ncol=4) # population covariance matrix

# generate sample means and covariance matrices.
# in each "means" matrix of sample means, each row is a simulation and
#                                         each column is a cell: 11, 12, 21, 22.

# in each "covs" array, each slice in the 3rd dimension is a 4x4 covariance matrix for
# a given simulation.

# if nSubj>4, generate covariance matrices directly from the Wishart distribution
# (see Anderson 1958, An Introduction to Multivariate Statistical Analysis, p. 159,
# Theorem 7.2.2) and generate means directly from multivariate normal distribution.

# if nSubj<5 , generate random observations from multivariate normal distribution:
# the observations start as a "temp1" matrix in which rows are subjects and columns
# are cells, then are reshaped into a "temp2" array in which each slice is a 4-by-n
# matrix for given simulation, then are reshaped into a "y" array in which each slice
# is an n-by-4 matrix for a given simulation. then the means matrix and covs array can
# be computed from the observations.

if (nSubj > 4) {
  means = mvrnorm( numSim, mu  , popCov/nSubj)  # means matrix
  covs  = rWishart(numSim, degf, popCov) / degf # covs matrix

} else {
  yTemp1 = mvrnorm(numSim*nSubj, mu, popCov)          # temp1 matrix
  yTemp2 = array(t(yTemp1), c(4, nSubj, numSim))      # temp2 array
  y = array(apply(yTemp2, 3, t), c(nSubj, 4, numSim)) # y array

  means = apply(y, c(3, 2), mean)                  # means matrix
  covs  = array(apply(y, 3, cov), c(4, 4, numSim)) # covs array
}

# define contrasts
contrastInter = c(1, -1, -1,  1)
contrastMain1 = c(1,  1, -1, -1)
contrastMain2 = c(1, -1,  1, -1)
contrast11v12 = c(1, -1,  0,  0)
contrast11v21 = c(1,  0, -1,  0)
contrast22v12 = c(0, -1,  0,  1)
contrast22v21 = c(0,  0, -1,  1)

# compute means for contrasts
meanInter = means %*% contrastInter
meanMain1 = means %*% contrastMain1
meanMain2 = means %*% contrastMain2
mean11v12 = means %*% contrast11v12
mean11v21 = means %*% contrast11v21
mean22v12 = means %*% contrast22v12
mean22v21 = means %*% contrast22v21

# define functions for computing quadratic forms from given covariance matrix
quadFormInter = function(covX) t(contrastInter) %*% covX %*% contrastInter
quadFormMain1 = function(covX) t(contrastMain1) %*% covX %*% contrastMain1
quadFormMain2 = function(covX) t(contrastMain2) %*% covX %*% contrastMain2
quadForm11v12 = function(covX) t(contrast11v12) %*% covX %*% contrast11v12
quadForm11v21 = function(covX) t(contrast11v21) %*% covX %*% contrast11v21
quadForm22v12 = function(covX) t(contrast22v12) %*% covX %*% contrast22v12
quadForm22v21 = function(covX) t(contrast22v21) %*% covX %*% contrast22v21
```

```
# compute standard errors for contrasts
semInter = sqrt(apply(covs, 3, quadFormInter))
semMain1 = sqrt(apply(covs, 3, quadFormMain1))
semMain2 = sqrt(apply(covs, 3, quadFormMain2))
sem11v12 = sqrt(apply(covs, 3, quadForm11v12))
sem11v21 = sqrt(apply(covs, 3, quadForm11v21))
sem22v12 = sqrt(apply(covs, 3, quadForm22v12))
sem22v21 = sqrt(apply(covs, 3, quadForm22v21))

# absolute-value t-statistics
absTInter = abs(sqrt(nSubj) * meanInter / semInter)
absTMain1 = abs(sqrt(nSubj) * meanMain1 / semMain1)
absTMain2 = abs(sqrt(nSubj) * meanMain2 / semMain2)
absT11v12 = abs(sqrt(nSubj) * mean11v12 / sem11v12)
absT11v21 = abs(sqrt(nSubj) * mean11v21 / sem11v21)
absT22v12 = abs(sqrt(nSubj) * mean22v12 / sem22v12)
absT22v21 = abs(sqrt(nSubj) * mean22v21 / sem22v21)

# test the interaction
cpa      = -qt(a/2, degf)  # critical point for two-sided testing at level a
sigInter = absTInter > cpa # flag significant interaction tests

# p-values for simple-effect t-statistics
p11v12 = 2*pt(-absT11v12, degf)
p11v21 = 2*pt(-absT11v21, degf)
p22v12 = 2*pt(-absT22v12, degf)
p22v21 = 2*pt(-absT22v21, degf)

# matrix of p-values (each row is for a given simulation)
pMat = cbind(p11v12, p11v21, p22v12, p22v21)

# null hypothesis statuses
nullMain1 = (mu11+mu12 == mu21+mu22)
nullMain2 = (mu11+mu21 == mu12+mu22)
null11v12 = (mu11 == mu12)
null11v21 = (mu11 == mu21)
null22v12 = (mu22 == mu12)
null22v21 = (mu22 == mu21)

# number of simple false nulls
numSimpleFalseNull = sum(!null11v12, !null11v21, !null22v12, !null22v21)

# are all nulls true?
nullAll = (numSimpleFalseNull == 0)

# flag significant tests for one-track AC method
sig11v12AC1 = (p11v12 < aComp1)
sig11v21AC1 = (p11v21 < aComp1)
sig22v12AC1 = (p22v12 < aComp1)
sig22v21AC1 = (p22v21 < aComp1)

# flag significant tests for two-track AC method
cpaComp2 = -qt(aComp2/2, degf) # critical point for two-sided testing at level aComp2

sigMain1AC2 = (absTMain1 > cpaComp2) & !sigInter
sigMain2AC2 = (absTMain2 > cpaComp2) & !sigInter

sig11v12AC2 = (p11v12 < aComp2) & sigInter
sig11v21AC2 = (p11v21 < aComp2) & sigInter
sig22v12AC2 = (p22v12 < aComp2) & sigInter
sig22v21AC2 = (p22v21 < aComp2) & sigInter
```

```
# flag significant tests for two-track method with no adjustment
sigMain1NoAdj = (absTMain1 > cpa) & !sigInter
sigMain2NoAdj = (absTMain2 > cpa) & !sigInter

sig11v12NoAdj = (p11v12 < a) & sigInter
sig11v21NoAdj = (p11v21 < a) & sigInter
sig22v12NoAdj = (p22v12 < a) & sigInter
sig22v21NoAdj = (p22v21 < a) & sigInter

# flag significant tests for two-track "given-track" Bonferroni method
# (adjusts for tests on reached track only)
cpaBonfMain = -qt(a/4, degf) # critical point for two-sided testing at level alpha/2
cpaBonfSimp = -qt(a/8, degf) # critical point for two-sided testing at level alpha/4

sigMain1Bonf = (absTMain1 > cpaBonfMain) & !sigInter
sigMain2Bonf = (absTMain2 > cpaBonfMain) & !sigInter
sig11v12Bonf = (absT11v12 > cpaBonfSimp) &  sigInter
sig11v21Bonf = (absT11v21 > cpaBonfSimp) &  sigInter
sig22v12Bonf = (absT22v12 > cpaBonfSimp) &  sigInter
sig22v21Bonf = (absT22v21 > cpaBonfSimp) &  sigInter

# flag significant tests for one-track Hommel method
hommelAdjust = function(p) p.adjust(p, "hom") # function to do Hommel adjustments
pHom         = apply(pMat, 1, hommelAdjust)   # matrix of Hommel-adjusted p-values
                                              # (each column is for a given sim)

sig11v12Hom = pHom[1,] < aHom
sig11v21Hom = pHom[2,] < aHom
sig22v12Hom = pHom[3,] < aHom
sig22v21Hom = pHom[4,] < aHom

# EWER for one-track AC method
anySigAC1TrueNull = (sig11v12AC1 & null11v12) | (sig11v21AC1 & null11v21) |
                    (sig22v12AC1 & null22v12) | (sig22v21AC1 & null22v21)

ewerAC1 = mean(anySigAC1TrueNull)

# EWER for two-track AC method
anySigAC2TrueNull = (sigMain1AC2 & nullMain1) | (sigMain2AC2 & nullMain2) |
                    (sig11v12AC2 & null11v12) | (sig11v21AC2 & null11v21) |
                    (sig22v12AC2 & null22v12) | (sig22v21AC2 & null22v21)

ewerAC2 = mean(anySigAC2TrueNull)

# EWER for two-track approach with no adjustment
anySigNoAdjTrueNull = (sigMain1NoAdj & nullMain1) | (sigMain2NoAdj & nullMain2) |
                      (sig11v12NoAdj & null11v12) | (sig11v21NoAdj & null11v21) |
                      (sig22v12NoAdj & null22v12) | (sig22v21NoAdj & null22v21)

ewerNoAdj = mean(anySigNoAdjTrueNull)

# EWER for two-track "given-track" Bonferroni method
# (adjusts for tests on reached track only)
anySigBonfTrueNullGiven = (sigMain1Bonf & nullMain1) | (sigMain2Bonf & nullMain2) |
                          (sig11v12Bonf & null11v12) | (sig11v21Bonf & null11v21) |
                          (sig22v12Bonf & null22v12) | (sig22v21Bonf & null22v21)

ewerBonfGiven = mean(anySigBonfTrueNullGiven)
```

```
# EWER for two-track "partial" Bonferroni method (only adjusts on simple-effect track)
anySigBonfTrueNullPart = (sigMain1NoAdj & nullMain1) | (sigMain2NoAdj & nullMain2) |
                         (sig11v12Bonf  & null11v12) | (sig11v21Bonf  & null11v21) |
                         (sig22v12Bonf  & null22v12) | (sig22v21Bonf  & null22v21)

ewerBonfPart = mean(anySigBonfTrueNullPart)

# "de facto" EWER for two-track AC method (considers main effects as Type I errors if
# null for either pairwise test of given factor is true)
anySigACTrueNullDF = anySigAC2TrueNull | (sigMain1AC2 & (null11v21 | null22v12)) |
                                         (sigMain2AC2 & (null11v12 | null22v21))

ewerACDF = mean(anySigACTrueNullDF)

# EWER for one-track Hommel method
anySigHomTrueNull = (sig11v12Hom & null11v12) | (sig11v21Hom & null11v21) |
                    (sig22v12Hom & null22v12) | (sig22v21Hom & null22v21)

ewerHom = mean(anySigHomTrueNull)

# compute power
if (nullAll) {
  avgPowerAC1 = NaN
  avgPowerHom = NaN
  anyPowerAC1 = NaN
  anyPowerAC2 = NaN
  anyPowerHom = NaN

} else {

  # per-test power for one-track AC method
  numSigAC1FalseNull = (sig11v12AC1 & !null11v12) + (sig11v21AC1 & !null11v21) +
                       (sig22v12AC1 & !null22v12) + (sig22v21AC1 & !null22v21)

  avgPowerAC1 = mean(numSigAC1FalseNull / numSimpleFalseNull)

  # any-test power for one-track AC method
  anySigAC1FalseNull = numSigAC1FalseNull > 0
  anyPowerAC1        = mean(anySigAC1FalseNull)

  # any-test power for two-track AC method
  anySigAC2FalseNull = (sigMain1AC2 & !nullMain1) | (sigMain2AC2 & !nullMain2) |
                       (sig11v12AC2 & !null11v12) | (sig11v21AC2 & !null11v21) |
                       (sig22v12AC2 & !null22v12) | (sig22v21AC2 & !null22v21)

  anyPowerAC2 = mean(anySigAC2FalseNull)

  # per-test power for one-track Hommel method
  numSigHomFalseNull = (sig11v12Hom & !null11v12) + (sig11v21Hom & !null11v21) +
                       (sig22v12Hom & !null22v12) + (sig22v21Hom & !null22v21)

  avgPowerHom = mean(numSigHomFalseNull / numSimpleFalseNull)

  # any-test power for one-track Hommel method
  anySigHomFalseNull = numSigHomFalseNull > 0
  anyPowerHom        = mean(anySigHomFalseNull)
}

totalTime = (proc.time()[3] - startTime) / 60 # total time elapsed in minutes

################################################################################
```

```
# REPORT PARAMETERS AND RESULTS

totalTime

numSim # number of simulations
a      # experimentwise alpha level
aComp1 # comparisonwise alpha level for one-track AC method
aComp2 # comparisonwise alpha level for two-track AC method
aHom   # nominal familywise alpha level for one-track Hommel method
nSubj  # number of subjects
mu     # population cell means
rho    # population covariance

avgPowerAC1 # per-test power for the one-track aComp method
avgPowerHom # per-test power for the one-track Hommel method

anyPowerAC1 # any-test power for the one-track aComp method
anyPowerAC2 # any-test power for the two-track aComp method
anyPowerHom # any-test power for one-track Hommel method

ewerAC1       # experimentwise Type I error rate for the one-track aComp method
ewerAC2       # experimentwise Type I error rate for the two-track aComp method
ewerNoAdj     # experimentwise Type I error rate for the two-track no-adjustment
ewerBonfGiven # experimentwise Type I error rate for the two-track given-track Bonfer.
ewerBonfPart  # experimentwise Type I error rate for the two-track partial Bonferroni
ewerACDF      # "de facto" experimentwise Type I error rate for the two-track aComp
ewerHom       # experimentwise Type I error rate for one-track Hommel method
```

APPENDIX L:

R CODE FOR `sim2x2mixed.R`

```
# These simulations compare any-test power and experimentwise error rates (EWERs) of
# different methods in a 2x2 mixed design. The "two-track" methods test only the main
# effects when the interaction is nonsignificant, and test only the simple (pairwise)
# effects when the interaction is significant. The "one-track" methods test only the
# simple-effect tests (without considering the interaction or main effects).

# Author: Andrew V. Frane (3/2019 using R version 3.3.3)

rm(list=ls())              # clear variables from workspace
startTime = proc.time()[3] # start timing this program

# load 'MASS' package (for mvrnorm function)
if(!require(MASS)) {       # try to load package; and if it isn't found...
  install.packages("MASS") # ...install it...
  require(MASS)            # ...and load it
}

################################################################################

# INPUT PARAMETERS

numSim    = 10^6  # number of simulations
a         = .05   # experimentwise alpha level

aComp1    = .0138 # comparisonwise alpha level for one-track AC method;
                  # obtain using 'ac2x2mixedsimp.R'

aComp2    = .0149 # comparisonwise alpha level for two-track AC method;
                  # obtain using 'ac2x2mixed.R'

aHom      = .0535 # nominal familywise alpha level for one-track Hommel method;
                  # obtain using 'ac2x2mixedsimp.R'

rho       = 0     # population covariance for within-subjects factor
                  # (can be 0 if nonnegativity assumed)

nGroup1   = 10    # number of subjects in group 1
nGroup2   = 10    # number of subjects in group 2

mu11      = 0     # standardized pop mean for level 1 of wthn-subjs factor in group 1
mu12      = 0     # standardized pop mean for level 2 of wthn-subjs factor in group 1
mu21      = 0     # standardized pop mean for level 1 of wthn-subjs factor in group 2
mu22      = 0     # standardized pop mean for level 2 of wthn-subjs factor in group 2

poolIntSS = TRUE  # whether interaction is dropped from model when testing on the
                  # main-effect track, i.e., whether to pool the interaction SS with
                  # within-cell SS & add 1 df for the within-subjects main-effect test

################################################################################
```

```
# SIMULATIONS

# define covariance matrices
popCov1112 = matrix(c( 1 ,rho,
                       rho, 1), ncol=2) # population covariance matrix for group 1

popCov2122 = matrix(c( 1, rho,
                       rho, 1), ncol=2) # population covariance matrix for group 2

# define degrees of freedom
degf1 = nGroup1 - 1
degf2 = nGroup2 - 1
degf  = degf1   + degf2
degfx = degf    + poolIntSS

# generate sample means and covariance matrices:
# in each "means" matrix of sample means, each row is a simulation and
#                                         each column is a cell: 11, 12, 21, 22.

# in each "covs" array, each slice in the 3rd dimension is a 2x2 covariance matrix for
# a given simulation.

# if n>2 for the given group, generate covariance matrices directly from Wishart
# distribution (see Anderson 1958, An Introduction to Multivariate Statistical
# Analysis, p. 159, Theorem 7.2.2) and generate means directly from the multivariate
# normal distribution.

# if n<3 for the given group, generate random observations from multivariate normal
# distribution: the observations start as a "temp1" matrix in which rows are subjects
# and columns are cells, then are reshaped into a "temp2" array in which each slice is
# a 2-by-n matrix for given simulation, then are reshaped into a "y" array in which
# each slice is an n-by-2 matrix for a given simulation. then the means matrix and
# covs array can be computed from the observations.

if (nGroup1 > 2) {

  means1112 = mvrnorm(numSim, c(mu11, mu12), popCov1112 / nGroup1)    # means mtrx grp1
  covs1112  = rWishart(numSim, degf1, popCov1112) / degf1             # covs array grp1

} else {

  y1112Temp1 = mvrnorm(numSim*nGroup1, c(mu11, mu12), popCov1112)     # temp1 mtrx grp1
  y1112Temp2 = array(t(y1112Temp1), c(2, nGroup1, numSim))            # temp2 arry grp1
  y1112      = array(apply(y1112Temp2, 3, t), c(nGroup1, 2, numSim)) # y array group 1
  means1112  = apply(y1112, c(3, 2), mean)                           # means mtrx grp1
  covs1112   = array(apply(y1112, 3, cov), c(2, 2, numSim))          # covs array grp1
  }

if (nGroup2 > 2) {
  means2122 = mvrnorm(numSim, c(mu21, mu22), popCov2122 / nGroup2)    # means mtrx grp2
  covs2122  = rWishart(numSim, degf2, popCov2122) / degf2             # covs array grp2

} else {
  y2122Temp1 = mvrnorm(numSim*nGroup2, c(mu21, mu22), popCov2122)     # temp1 mtrx grp2
  y2122Temp2 = array(t(y2122Temp1), c(2, nGroup2, numSim))            # temp2 arry grp2
  y2122      = array(apply(y2122Temp2, 3, t), c(nGroup2, 2, numSim)) # y array group 2
  means2122  = apply(y2122, c(3, 2), mean)                           # means mtrx grp2
  covs2122   = array(apply(y2122, 3, cov), c(2, 2, numSim))          # covs array grp2
  }

# sample pooled within-group 2x2 covariance matrix for repeated measures
covsPool = (degf1*covs1112 + degf2*covs2122) / degf
```

```
# define contrasts for mean comparisons
nAvg = (nGroup1 + nGroup2) / 2
f    =  nGroup1 / nAvg
g    = nGroup2 / nAvg

contrastInter = c( 1, -1, -1,  1) # interaction
contrastMainB = c( 1,  1, -1, -1) # between-groups main effect
contrastMainW = c( f, -f,  g, -g) # within-groups  main effect weighted by group sizes
contrast11v12 = c( 1, -1)         # simple effect within group 1
contrast22v21 = c(-1,  1)         # simple effect within group 2
contrast11v21 = c( 1,  0, -1,  0) # between-groups simple effect for first measure
contrast22v12 = c( 0, -1,  0,  1) # between-groups simple effect for second measure

# compute sample mean differences
means     = cbind(means1112, means2122)
meanInter = means     %*% contrastInter
meanMainB = means     %*% contrastMainB
meanMainW = means     %*% contrastMainW
mean11v21 = means     %*% contrast11v21
mean22v12 = means     %*% contrast22v12
mean11v12 = means1112 %*% contrast11v12
mean22v21 = means2122 %*% contrast22v21

adjOne = sqrt(nGroup1+nGroup2) / 2
adjTwo = sqrt(nGroup1*nGroup2  / (nGroup1+nGroup2))
adjGr1 = sqrt(nGroup1)
adjGr2 = sqrt(nGroup2)

# contrasts for sample covariance matrices
contrastDif = c(1, -1)
contrastSum = c(1,  1)

# define quadratic forms used to compute variances
quadFormSum   = function(covX) t(contrastSum)   %*% covX %*% contrastSum
quadFormDif   = function(covX) t(contrastDif)   %*% covX %*% contrastDif
quadForm11v12 = function(covX) t(contrast11v12) %*% covX %*% contrast11v12
quadForm22v21 = function(covX) t(contrast22v21) %*% covX %*% contrast22v21
quadFormVar1  = function(covX) covX[1, 1]
quadFormVar2  = function(covX) covX[2, 2]

# compute variances
varDif   = apply(covsPool, 3, quadFormDif)
varMainB = apply(covsPool, 3, quadFormSum)
var11v21 = apply(covsPool, 3, quadFormVar1)
var22v12 = apply(covsPool, 3, quadFormVar2)
var11v12 = apply(covs1112, 3, quadForm11v12)
var22v21 = apply(covs2122, 3, quadForm22v21)

varMainW = (varDif*degf + ((adjOne*meanInter)^2)*poolIntSS) / degfx

# compute t-statistics
tInter = adjTwo * meanInter / sqrt(varDif  ) # two-sample t-test of interaction
tMainB = adjTwo * meanMainB / sqrt(varMainB) # two-sample t-test btwn-grp main effect
tMainW = adjOne * meanMainW / sqrt(varMainW) # one-sample t-test wthn-grp main effect;
                                             # df = degfx

t11v12 = adjGr1 * mean11v12 / sqrt(var11v12) # one-sample t-test within group 1
t22v21 = adjGr2 * mean22v21 / sqrt(var22v21) # one-sample t-test within group 2
t11v21 = adjTwo * mean11v21 / sqrt(var11v21) # two-sample t-test between groups
                                             # at level 1 of within-group factor
```

```
t22v12 = adjTwo * mean22v12 / sqrt(var22v12) # two-sample t-test between groups
                                             # at level 2 of within-group factor

# absolute value of t-statistics
absTInter = abs(tInter)
absTMainB = abs(tMainB)
absTMainW = abs(tMainW)
absT11v12 = abs(t11v12)
absT11v21 = abs(t11v21)
absT22v12 = abs(t22v12)
absT22v21 = abs(t22v21)

# test the interaction
cpadegf  = -qt(a/2, degf)      # unadjusted crit point for 2-sided test with df=degf
sigInter = absTInter > cpadegf # flag significant interaction tests

# p-values for simple-effect t-statistics
p11v12 = 2*pt(-absT11v12, degf1)
p11v21 = 2*pt(-absT11v21, degf )
p22v12 = 2*pt(-absT22v12, degf )
p22v21 = 2*pt(-absT22v21, degf2)

# matrix of p-values (each row is for a given simulation)
pMat = cbind(p11v12, p11v21, p22v12, p22v21)

# null hypothesis statuses
nullMainB = (mu11+mu12 == mu21+mu22)
nullMainW = (mu11+mu21 == mu12+mu22)
null11v12 = (mu11 == mu12)
null11v21 = (mu11 == mu21)
null22v12 = (mu22 == mu12)
null22v21 = (mu22 == mu21)

# number of simple false nulls
numSimpleFalseNull = sum(!null11v12, !null11v21, !null22v12, !null22v21)

# are all nulls true?
nullAll = (numSimpleFalseNull == 0)

# flag significant tests for one-track AC method
sig11v12AC1 = (p11v12 < aComp1)
sig11v21AC1 = (p11v21 < aComp1)
sig22v12AC1 = (p22v12 < aComp1)
sig22v21AC1 = (p22v21 < aComp1)

# flag significant tests for two-track AC method
cpaComp2degf  = -qt(aComp2/2, degf ) # crit point for btwn-grp main-effect at aComp2
cpaComp2degfx = -qt(aComp2/2, degfx) # crit point for wthn-grp main effect at aComp2

sigMainBAC2 = (absTMainB > cpaComp2degf ) & !sigInter
sigMainWAC2 = (absTMainW > cpaComp2degfx) & !sigInter

sig11v12AC2 = (p11v12 < aComp2) & sigInter
sig11v21AC2 = (p11v21 < aComp2) & sigInter
sig22v12AC2 = (p22v12 < aComp2) & sigInter
sig22v21AC2 = (p22v21 < aComp2) & sigInter

# flag significant tests for two-track method with no adjustment
cpadegfx = -qt(a/2, degfx) # crit point for within-grp main effect at unadjusted alpha

sigMainBNoAdj = (absTMainB > cpadegf ) & !sigInter
sigMainWNoAdj = (absTMainW > cpadegfx) & !sigInter
```

```
sig11v12NoAdj = (p11v12 < a) & sigInter
sig11v21NoAdj = (p11v21 < a) & sigInter
sig22v12NoAdj = (p22v12 < a) & sigInter
sig22v21NoAdj = (p22v21 < a) & sigInter

# flag significant tests for 2-track "given-track" Bonferroni
# (adjusts for tests on reached track only)
cpaBonfMainB = -qt(a/4, degf ) # critical point for between-group main effect at a/2
cpaBonfMainW = -qt(a/4, degfx) # critical point for within- group main effect at a/2
cpaBonf11v12 = -qt(a/8, degf1) # critical point for two-sided test of 11v12 at a/4
cpaBonf11v21 = -qt(a/8, degf ) # critical point for two-sided test of 11v21 at a/4
cpaBonf22v12 = -qt(a/8, degf ) # critical point for two-sided test of 22v12 at a/4
cpaBonf22v21 = -qt(a/8, degf2) # critical point for two-sided test of 22v21 at a/4

sigMainBBonf = (absTMainB > cpaBonfMainB) & !sigInter
sigMainWBonf = (absTMainW > cpaBonfMainW) & !sigInter
sig11v12Bonf = (absT11v12 > cpaBonf11v12) &  sigInter
sig11v21Bonf = (absT11v21 > cpaBonf11v21) &  sigInter
sig22v12Bonf = (absT22v12 > cpaBonf22v12) &  sigInter
sig22v21Bonf = (absT22v21 > cpaBonf22v21) &  sigInter

# flag significant tests for one-track Hommel method
hommelAdjust = function(p) p.adjust(p, "hom") # function to do Hommel adjustments
pHom         = apply(pMat, 1, hommelAdjust)   # matrix of Hommel-adjusted p-values
                                              # (each column is for a given sim)

sig11v12Hom = pHom[1,] < aHom
sig11v21Hom = pHom[2,] < aHom
sig22v12Hom = pHom[3,] < aHom
sig22v21Hom = pHom[4,] < aHom

# EWER for one-track AC method
anySigAC1TrueNull = (sig11v12AC1 & null11v12) | (sig11v21AC1 & null11v21) |
                    (sig22v12AC1 & null22v12) | (sig22v21AC1 & null22v21)

ewerAC1 = mean(anySigAC1TrueNull)

# EWER for two-track AC method
anySigAC2TrueNull = (sigMainBAC2 & nullMainB) | (sigMainWAC2 & nullMainW) |
                    (sig11v12AC2 & null11v12) | (sig11v21AC2 & null11v21) |
                    (sig22v12AC2 & null22v12) | (sig22v21AC2 & null22v21)

ewerAC2 = mean(anySigAC2TrueNull)

# EWER for two-track approach with no adjustment
anySigNoAdjTrueNull = (sigMainBNoAdj & nullMainB) | (sigMainWNoAdj & nullMainW) |
                      (sig11v12NoAdj & null11v12) | (sig11v21NoAdj & null11v21) |
                      (sig22v12NoAdj & null22v12) | (sig22v21NoAdj & null22v21)

ewerNoAdj = mean(anySigNoAdjTrueNull)

# EWER for two-track "given-track" Bonferroni method
# (adjusts for tests on reached track only)
anySigBonfTrueNullGiven = (sigMainBBonf & nullMainB) | (sigMainWBonf & nullMainW) |
                          (sig11v12Bonf & null11v12) | (sig11v21Bonf & null11v21) |
                          (sig22v12Bonf & null22v12) | (sig22v21Bonf & null22v21)

ewerBonfGiven = mean(anySigBonfTrueNullGiven)

# EWER for two-track "partial" Bonferroni method (only adjusts on simple-effect track)
anySigBonfTrueNullPart = (sigMainBNoAdj & nullMainB) | (sigMainWNoAdj & nullMainW) |
```

```
                          (sig11v12Bonf  & null11v12) | (sig11v21Bonf  & null11v21) |
                          (sig22v12Bonf  & null22v12) | (sig22v21Bonf  & null22v21)

ewerBonfPart = mean(anySigBonfTrueNullPart)

# "de facto" EWER for two-track AC method (considers main effects as Type I errors if
# null for either pairwise test of given factor is true)
anySigACTrueNullDF = anySigAC2TrueNull | (sigMainBAC2 & (null11v21 | null22v12)) |
                                         (sigMainWAC2 & (null11v12 | null22v21))

ewerACDF = mean(anySigACTrueNullDF)

# EWER for one-track Hommel method
anySigHomTrueNull = (sig11v12Hom & null11v12) | (sig11v21Hom & null11v21) |
                    (sig22v12Hom & null22v12) | (sig22v21Hom & null22v21)

ewerHom = mean(anySigHomTrueNull)

# compute power
if (nullAll) {
  avgPowerAC1 = NaN
  avgPowerHom = NaN
  anyPowerAC1 = NaN
  anyPowerAC2 = NaN
  anyPowerHom = NaN

} else {

  # per-test power for one-track AC method
  numSigAC1FalseNull = (sig11v12AC1 & !null11v12) + (sig11v21AC1 & !null11v21) +
                       (sig22v12AC1 & !null22v12) + (sig22v21AC1 & !null22v21)

  avgPowerAC1 = mean(numSigAC1FalseNull / numSimpleFalseNull)

  # any-test power for one-track AC method
  anySigAC1FalseNull = numSigAC1FalseNull > 0
  anyPowerAC1        = mean(anySigAC1FalseNull)

  # any-test power for two-track AC method
  anySigAC2FalseNull = (sigMainBAC2 & !nullMainB) | (sigMainWAC2 & !nullMainW) |
                       (sig11v12AC2 & !null11v12) | (sig11v21AC2 & !null11v21) |
                       (sig22v12AC2 & !null22v12) | (sig22v21AC2 & !null22v21)

  anyPowerAC2 = mean(anySigAC2FalseNull)

  # per-test power for one-track Hommel method
  numSigHomFalseNull = (sig11v12Hom & !null11v12) + (sig11v21Hom & !null11v21) +
                       (sig22v12Hom & !null22v12) + (sig22v21Hom & !null22v21)

  avgPowerHom = mean(numSigHomFalseNull / numSimpleFalseNull)

  # any-test power for one-track Hommel method
  anySigHomFalseNull = numSigHomFalseNull > 0
  anyPowerHom        = mean(anySigHomFalseNull)
}

totalTime = (proc.time()[3] - startTime) / 60 # total time elapsed in minutes

################################################################################
```

```
# REPORT PARAMETERS AND RESULTS

totalTime

numSim                    # number of simulations
a                         # experimentwise alpha level
aComp1                    # comparisonwise alpha level for one-track AC method
aComp2                    # comparisonwise alpha level for two-track AC method
aHom                      # nominal familywise alpha level for one-track Hommel method
c(nGroup1, nGroup2)       # number of subjects in each group
c(mu11, mu12, mu21, mu22) # standardized population cell means
rho                       # population covariance for within-subjects factor
poolIntSS                 # interaction dropped from model on main-effect track?

avgPowerAC1 # per-test power for the one-track aComp method
avgPowerHom # per-test power for the one-track Hommel method

anyPowerAC1 # any-test power for the one-track aComp method
anyPowerAC2 # any-test power for the two-track aComp method
anyPowerHom # any-test power for Hommel-adjusted simple-effect tests

ewerAC1       # experimentwise Type I error rate for the one-track aComp method
ewerAC2       # experimentwise Type I error rate for the two-track aComp method
ewerNoAdj     # experimentwise Type I error rate for the two-track no-adjustment
ewerBonfGiven # experimentwise Type I error rate for the two-track given-track Bonfer.
ewerBonfPart  # experimentwise Type I error rate for the two-track partial Bonferroni
ewerACDF      # "de facto" experimentwise Type I error rate for two-track aComp method
ewerHom       # experimentwise Type I error rate for the one-track Hommel method
```

APPENDIX M:

R CODE FOR `ac2x2betweensimp.R`

```
# These simulations find the optimal uniform comparisonwise alpha level (aComp) for
# the simple-effect tests in a 2x2 between-subjects design in order to control the
# experimentwise Type I error rate at the desired alpha level. The simulations also
# find the optimal nominal familywise alpha level (aHom) for the Hommel procedure as
# applied to the simple-effects tests.

# Author: Andrew V. Frane (4/2019 using R version 3.3.3)

rm(list=ls())              # clear variables from workspace
startTime = proc.time()[3] # start timing this program

################################################################################

# INPUT PARAMETERS

numSim = 10^7 # number of simulations
a      = .05  # familywise alpha level
n11    = 1000 # number of observations in cell 11
n12    = 1000 # number of observations in cell 12
n21    = 1000 # number of observations in cell 21
n22    = 1000 # number of observations in cell 22

################################################################################

# FIXED PARAMETERS
# (can be changed to explore scenarios of unequal population means and/or variances)

mu11    = 0 # population mean for cell 11
mu12    = 0 # population mean for cell 12
mu21    = 0 # population mean for cell 21
mu22    = 0 # population mean for cell 22

sigma11 = 1 # population standard deviation for cell 11
sigma12 = 1 # population standard deviation for cell 12
sigma21 = 1 # population standard deviation for cell 21
sigma22 = 1 # population standard deviation for cell 22

################################################################################

# SIMULATIONS

# degrees of freedom
degf11 = n11 - 1
degf12 = n12 - 1
degf21 = n21 - 1
degf22 = n22 - 1

degf11v12 = degf11 + degf12
degf11v21 = degf11 + degf21
degf22v12 = degf22 + degf12
degf22v21 = degf22 + degf21
```

```
# randomly generate sample means from normal distribution
means11 = rnorm(numSim, mu11, sigma11/sqrt(n11))
means12 = rnorm(numSim, mu12, sigma12/sqrt(n12))
means21 = rnorm(numSim, mu21, sigma21/sqrt(n21))
means22 = rnorm(numSim, mu22, sigma22/sqrt(n22))

# contrasts
mean11v12 = means11 - means12
mean11v21 = means11 - means21
mean22v12 = means22 - means12
mean22v21 = means22 - means21

# sums of squares
ss11 = rchisq(numSim, degf11)
ss12 = rchisq(numSim, degf12)
ss21 = rchisq(numSim, degf21)
ss22 = rchisq(numSim, degf22)

# standard deviations
sd11v12 = sqrt((ss11+ss12) / degf11v12)
sd11v21 = sqrt((ss11+ss21) / degf11v21)
sd22v12 = sqrt((ss22+ss12) / degf22v12)
sd22v21 = sqrt((ss22+ss21) / degf22v21)

# coefficients used to adjust numerator when computing t-statistics
adj11v12 = 1/sqrt(1/n11 + 1/n12)
adj11v21 = 1/sqrt(1/n11 + 1/n21)
adj22v12 = 1/sqrt(1/n22 + 1/n12)
adj22v21 = 1/sqrt(1/n22 + 1/n21)

# t-tests
t11v12 = adj11v12 * mean11v12 / sd11v12
t11v21 = adj11v21 * mean11v21 / sd11v21
t22v12 = adj22v12 * mean22v12 / sd22v12
t22v21 = adj22v21 * mean22v21 / sd22v21

# absolute-values of t-statistics
absT11v12 = abs(t11v12)
absT11v21 = abs(t11v21)
absT22v12 = abs(t22v12)
absT22v21 = abs(t22v21)

# raw p-values for simple-effect t-statistics
p11v12 = 2*pt(-absT11v12, degf11v12)
p11v21 = 2*pt(-absT11v21, degf11v21)
p22v12 = 2*pt(-absT22v12, degf22v12)
p22v21 = 2*pt(-absT22v21, degf22v21)

# matrix of p-values (each row is for a given simulation)
pMat = cbind(p11v12, p11v21, p22v12, p22v21)

# null hypothesis statuses
null11v12 =  mu11 == mu12
null11v21 =  mu11 == mu21
null22v12 =  mu12 == mu22
null22v21 =  mu21 == mu22

# minimum raw simple-effect p-value for a true null hypothesis
minPTrueNull = pmin(p11v12 + !null11v12, p11v21 + !null11v21,
                    p22v12 + !null22v12, p22v21 + !null22v21)
```

```
# minimum Hommel-adjusted simple-effect p-value for a true null hypothesis
hommelAdjust     = function(p) p.adjust(p, "hom")
minWhereTrueNull = function(p) min(p[c(null11v12, null11v21,
                                       null22v12, null22v21)==TRUE])

# matrix of Hommel-adjusted p-values (each column is for a given simulation)
pHom = apply(pMat, 1, hommelAdjust)

# minimum hommel-adjusted true-null p-value for each simulation
minPHomTrueNull = apply(pHom, 2, minWhereTrueNull)

################################################################################

# BINARY SEARCH FOR OPTIMAL UNIFORM COMPARISONWISE ALPHA LEVEL (aComp)
# FOR SIMPLE-EFFECT TESTS

targetNumSimsError = round(a*numSim) # target number of sims producing >=1 error

aComp               = a/4     # initialize comparisonwise alpha level
stepSize            = aComp/2 # initialize step size to nudge candidate aComp up & dwn
aCompFound          = FALSE   # initialize flag indicating if optimal cpaComp found
latestStepDirection = 0       # initialize latest step direction (0=down, 1=up)

while (aCompFound==FALSE) {

  # total number of simulations producing at least one Type I error
  numSimsError = sum(minPTrueNull < aComp)

  # nudge aComp up or down as necessary
  if (numSimsError > targetNumSimsError) {          # ERROR RATE IS TOO HIGH
    stepSize = stepSize / (latestStepDirection + 1) # if prev step up, halve step size
    aComp    = max(0, aComp - stepSize)             # nudge down aComp
    latestStepDirection = 0                         # set latest step direction to dwn

  } else if (numSimsError < targetNumSimsError) {   # ERROR RATE IS TOO LOW
    stepSize = stepSize / (2 - latestStepDirection) # if prev step dwn, halve stepsize
    aComp    = aComp + stepSize                     # nudge up aComp
    latestStepDirection = 1                         # set latest step direction to up

  } else {                                          # ERROR RATE IS JUST RIGHT
    aCompFound = TRUE                               # optimal aComp has been found
  }
}

################################################################################

# BINARY SEARCH FOR OPTIMAL NOMINAL FAMILYWISE ALPHA LEVEL (aHom)
# FOR SIMPLE-EFFECT TESTS IN THE ONE-TRACK HOMMEL PROCEDURE

targetNumSimsError = round(a*numSim) # target number of sims producing >=1 error
aHom               = a               # initialize nominal familywise Hommel alpha
stepSize           = a/10            # initialize step size to nudge aHom up and down

aHomFound           = FALSE # initialize flag indicating whether optimal aHom found
latestStepDirection = 0     # initialize latest step direction (0=down, 1=up)

while (aHomFound==FALSE) {

  numSimsError = sum(minPHomTrueNull < aHom) # total number of sims producing >=1 err
```

```
  # nudge aHom up or down as necessary
  if (numSimsError > targetNumSimsError) {          # ERROR RATE IS TOO HIGH
    stepSize = stepSize / (latestStepDirection + 1) # if prev step up, halve step size
    aHom     = max(0, aHom - stepSize)              # nudge down aHom
    latestStepDirection = 0                         # set latest step direction to dwn

  } else if (numSimsError < targetNumSimsError) {   # ERROR RATE IS TOO LOW
    stepSize = stepSize / (2 - latestStepDirection) # if prev step dwn, halve stepsize
    aHom     = aHom + stepSize                      # nudge up aHom
    latestStepDirection = 1                         # set latest step direction to up

  } else {                                          # ERROR RATE IS JUST RIGHT
    aHomFound = TRUE                                # optimal aHom has been found
  }
}

totalTime = (proc.time()[3] - startTime) / 60 # total time elapsed in minutes

################################################################################

# REPORT THE PARAMETERS AND RESULTS

totalTime

numSim                # number of simulations
a                     # experimentwise alpha level
c(n11, n12, n21, n22) # number of observations in each cell
aComp                 # optimal comparisonwise alpha level for one-track AC method
a/aComp               # ratio of experimentwise alpha level to aComp
aHom                  # nominal familywise alpha level for one-track Hommel method
```

APPENDIX N:

R CODE FOR `ac2x2withinsimp.R`

```
# These simulations find the optimal uniform comparisonwise alpha level (aComp) for
# the simple-effect tests in a 2x2 within-subjects design in order to control the
# experimentwise Type I error rate at the desired alpha level. The simulations also
# find the optimal nominal familywise alpha level (aHom) for the Hommel procedure as
# applied to the simple-effects tests.

# Author: Andrew V. Frane (4/2019 using R version 3.3.3)

rm(list=ls())              # clear variables from workspace
startTime = proc.time()[3] # start timing this program

# load 'MASS' package (for mvrnorm function)
if(!require(MASS)) {       # try to load package; and if it isn't found...
  install.packages("MASS") # ...install it...
  require(MASS)            # ...and load it
}

################################################################################

# INPUT PARAMETERS

numSim = 10^7 # number of simulations
a      = .05  # experimentwise alpha level
nSubj  = 100  # number of subjects

################################################################################

# FIXED PARAMETERS
# (can be changed to explore scenarios of unequal population means and/or variances)

mu11 = 0 # standardized population mean for cell 11
mu12 = 0 # standardized population mean for cell 12
mu21 = 0 # standardized population mean for cell 21
mu22 = 0 # standardized population mean for cell 22

popCov = matrix(c(1, 0, 0, 0,
                  0, 1, 0, 0,
                  0, 0, 1, 0,
                  0, 0, 0, 1), ncol=4) # population covariance matrix

################################################################################

# SIMULATIONS

mu   = c(mu11, mu12, mu21, mu22) # vector of population means
degf = nSubj - 1                 # degrees of freedom

# generate sample means and covariance matrices:
# in each "means" matrix of sample means, each row is a simulation and
#                                         each column is a cell: 11, 12, 21, 22.

# in each "covs" array, each slice in the 3rd dimension is a 4x4 covariance matrix for
# a given simulation.
```

```
# if nSubj>4, generate covariance matrices directly from the Wishart distribution (see
# Anderson 1958, An Introduction to Multivariate Statistical Analysis, p. 159, Theorem
# 7.2.2) and generate means directly from the multivariate normal distribution.

# if nSubj<5 , generate random observations from multivariate normal distribution: the
# observations start as a "temp1" matrix in which rows are subjects and columns are
# cells, then are reshaped into a "temp2" array in which each slice is a 4-by-n matrix
# for given simulation, then are reshaped into a "y" array in which each slice is an
# n-by-4 matrix for a given simulation. then the means matrix and covs array can be
# computed from the observations.

if (nSubj > 4) {
  means = mvrnorm( numSim, mu  , popCov/nSubj)  # means matrix
  covs  = rWishart(numSim, degf, popCov) / degf # covs matrix

} else {
  yTemp1 = mvrnorm(numSim*nSubj, mu, popCov)               # temp1 matrix
  yTemp2 = array(t(yTemp1), c(4, nSubj, numSim))           # temp2 array
  y      = array(apply(yTemp2, 3, t), c(nSubj, 4, numSim)) # y array

  means  = apply(y, c(3, 2), mean)                         # means matrix
  covs   = array(apply(y, 3, cov), c(4, 4, numSim))        # covs array
}

# define contrasts
contrast11v12 = c(1, -1,  0,  0)
contrast11v21 = c(1,  0, -1,  0)
contrast22v12 = c(0, -1,  0,  1)
contrast22v21 = c(0,  0, -1,  1)

# compute means for contrasts
mean11v12 = means %*% contrast11v12
mean11v21 = means %*% contrast11v21
mean22v12 = means %*% contrast22v12
mean22v21 = means %*% contrast22v21

# define functions for computing quadratic forms from given covariance matrix
quadForm11v12 = function(covX) t(contrast11v12) %*% covX %*% contrast11v12
quadForm11v21 = function(covX) t(contrast11v21) %*% covX %*% contrast11v21
quadForm22v12 = function(covX) t(contrast22v12) %*% covX %*% contrast22v12
quadForm22v21 = function(covX) t(contrast22v21) %*% covX %*% contrast22v21

# compute standard errors for contrasts
sem11v12 = sqrt(apply(covs, 3, quadForm11v12))
sem11v21 = sqrt(apply(covs, 3, quadForm11v21))
sem22v12 = sqrt(apply(covs, 3, quadForm22v12))
sem22v21 = sqrt(apply(covs, 3, quadForm22v21))

# compute t-statistics
t11v12 = sqrt(nSubj) * mean11v12 / sem11v12
t11v21 = sqrt(nSubj) * mean11v21 / sem11v21
t22v12 = sqrt(nSubj) * mean22v12 / sem22v12
t22v21 = sqrt(nSubj) * mean22v21 / sem22v21

# absolute value of t-statistics
absT11v12 = abs(t11v12)
absT11v21 = abs(t11v21)
absT22v12 = abs(t22v12)
absT22v21 = abs(t22v21)
```

```
# raw p-values for simple-effect t-statistics
p11v12 = 2*pt(-absT11v12, degf)
p11v21 = 2*pt(-absT11v21, degf)
p22v12 = 2*pt(-absT22v12, degf)
p22v21 = 2*pt(-absT22v21, degf)

# matrix of p-values (each row is for a given simulation)
pMat = cbind(p11v12, p11v21, p22v12, p22v21)

# null hypothesis statuses
null11v12  =  mu11 == mu12
null11v21  =  mu11 == mu21
null22v12  =  mu12 == mu22
null22v21  =  mu21 == mu22

# minimum raw p-value for a true null hypothesis
minPTrueNull = pmin(p11v12 + !null11v12, p11v21 + !null11v21,
                    p22v12 + !null22v12, p22v21 + !null22v21)

# minimum Hommel-adjusted simple-effect p-value for a true null hypothesis
hommelAdjust     = function(p) p.adjust(p, "hom")
minWhereTrueNull = function(p) min(p[c(null11v12, null11v21,
                                       null22v12, null22v21)==TRUE])

# matrix of Hommel-adjusted p-values (each column is for a given simulation)
pHom = apply(pMat, 1, hommelAdjust)

# minimum Hommel-adjusted true-null p-value for each simulation
minPHomTrueNull = apply(pHom, 2, minWhereTrueNull)

################################################################################

# BINARY SEARCH FOR OPTIMAL UNIFORM COMPARISONWISE ALPHA LEVEL (aComp)
# FOR SIMPLE-EFFECT TESTS

targetNumSimsError = round(a*numSim) # target number of sims producing >=1 error
aComp              = a/4             # initialize comparisonwise alpha level
stepSize           = aComp/2         # initialize step size to nudge aComp up & down

aCompFound          = FALSE # initialize flag indicating whether optimal cpaComp found
latestStepDirection = 0     # initialize latest step direction (0=down, 1=up)

while (aCompFound==FALSE) {

  # total number of simulations producing at least one Type I error
  numSimsError = sum(minPTrueNull < aComp)

  # nudge aComp up or down as necessary
  if (numSimsError > targetNumSimsError) {          # ERROR RATE IS TOO HIGH
    stepSize = stepSize / (latestStepDirection + 1) # if prev step up, halve step size
    aComp    = max(0, aComp - stepSize)             # nudge down aComp
    latestStepDirection = 0                         # set latest step direction to dwn

  } else if (numSimsError < targetNumSimsError) {   # ERROR RATE IS TOO LOW
    stepSize = stepSize / (2 - latestStepDirection) # if prev step dwn, halve stepsize
    aComp    = aComp + stepSize                     # nudge up aComp
    latestStepDirection = 1                         # set latest step direction to up

  } else {                                          # ERROR RATE IS JUST RIGHT
    aCompFound = TRUE                               # optimal aComp has been found
```

```
  }
}

################################################################################

# BINARY SEARCH FOR OPTIMAL NOMINAL FAMILYWISE ALPHA LEVEL (aHom)
# FOR SIMPLE-EFFECT TESTS IN THE ONE-TRACK HOMMEL PROCEDURE

targetNumSimsError  = round(a*numSim) # target number of sims producing >=1 error
aHom                = a               # initialize nominal familywise Hommel alpha
stepSize            = a/10            # initialize step size to nudge aHom up and down
aHomFound           = FALSE # initialize flag indicating whether optimal aHom found
latestStepDirection = 0     # initialize latest step direction (0=down, 1=up)

while (aHomFound==FALSE) {

  numSimsError = sum(minPHomTrueNull < aHom) # total number of sims with >=1 error

  # nudge aHom up or down as necessary
  if (numSimsError > targetNumSimsError) {          # ERROR RATE IS TOO HIGH
    stepSize = stepSize / (latestStepDirection + 1) # if prev step up, halve step size
    aHom     = max(0, aHom - stepSize)              # nudge down aHom
    latestStepDirection = 0                         # set latest step direction to dwn

  } else if (numSimsError < targetNumSimsError) {   # ERROR RATE IS TOO LOW
    stepSize = stepSize / (2 - latestStepDirection) # if prev step dwn, halve stepsize
    aHom     = aHom + stepSize                      # nudge up aHom
    latestStepDirection = 1                         # set latest step direction to up

  } else {                                          # ERROR RATE IS JUST RIGHT
    aHomFound = TRUE                                # optimal aHom has been found
  }
}

totalTime = (proc.time()[3] - startTime) / 60 # total time elapsed in minutes

################################################################################

# REPORT THE PARAMETERS AND RESULTS

totalTime

numSim  # number of simulations
a       # experimentwise alpha level
nSubj   # number of subjects
aComp   # optimal comparisonwise alpha level for one-track AC method
a/aComp # ratio of experimentwise alpha level to aComp
aHom    # nominal familywise alpha level for one-track Hommel method
```

APPENDIX O:

R CODE FOR `ac2x2mixedsimp.R`

```
# These simulations find the optimal uniform comparisonwise alpha level (aComp) for
# the simple-effect tests in a 2x2 mixed design in order to control the experimentwise
# Type I error rate at the desired alpha level. The simulations also find the optimal
# nominal familywise alpha level (aHom) for the Hommel procedure as applied to the
# simple-effect tests.

# Author: Andrew V. Frane (4/2019 using R version 3.3.3)

rm(list=ls())              # clear variables from workspace
startTime = proc.time()[3] # start timing this program

# load 'MASS' package (for mvrnorm function)
if(!require(MASS)) {       # try to load package; and if it isn't found...
  install.packages("MASS") # ...install it...
  require(MASS)            # ...and load it
}

################################################################################

# INPUT PARAMETERS

numSim  = 10^7 # number of simulations
a       = .05  # experimentwise alpha level
rho     = 0    # population covariance for within-subjects factor
               # (can be 0 if nonnegativity assumed)

nGroup1 = 1000 # number of subjects in group 1
nGroup2 = 1000 # number of subjects in group 2

################################################################################

# FIXED PARAMETERS (can be changed to explore scenarios of unequal population means
# and/or unequal covariance matrices)

mu11 = 0 # population mean for level 1 of within-subjects factor in group 1
mu12 = 0 # population mean for level 2 of within-subjects factor in group 1
mu21 = 0 # population mean for level 1 of within-subjects factor in group 2
mu22 = 0 # population mean for level 2 of within-subjects factor in group 2

popCov1112 = matrix(c( 1 , rho,
                       rho,  1), ncol=2) # population covariance matrix for group 1

popCov2122 = matrix(c( 1 , rho,
                       rho,  1), ncol=2) # population covariance matrix for group 2

################################################################################

# SIMULATIONS

# degrees of freedom
degf1 = nGroup1 - 1
degf2 = nGroup2 - 1
```

```
degf  = degf1   + degf2

# generate sample means and covariance matrices:
# in each "means" matrix of sample means, each row is a simulation and
#                                         each column is a cell: 11, 12, 21, 22.

# in each "covs" array, each slice in the 3rd dimension is a 2x2 covariance matrix for
# a given simulation.

# if n>2 for the given group, generate covariance matrices directly from the Wishart
# distribution (see Anderson 1958, An Introduction to Multivariate Statistical
# Analysis, p. 159, Theorem 7.2.2) and generate means directly from the multivariate
# normal distribution.

# if n<3 for the given group, generate random observations from multivariate normal
# distribution: the observations start as a "temp1" matrix in which rows are subjects
# and columns are cells, then are reshaped into a "temp2" array in which each slice is
# a 2-by-n matrix for given simulation, then are reshaped into a "y" array in which
# each slice is an n-by-2 matrix for a given simulation. then the means matrix and
# covs array can be computed from the observations.

if (nGroup1 > 2) {

  means1112 = mvrnorm(numSim, c(mu11, mu12), popCov1112 / nGroup1)    # means mtrx grp1
  covs1112  = rWishart(numSim, degf1, popCov1112) / degf1             # covs array grp1

} else {

  y1112Temp1 = mvrnorm(numSim*nGroup1, c(mu11, mu12), popCov1112)     # temp1 mtrx grp1
  y1112Temp2 = array(t(y1112Temp1), c(2, nGroup1, numSim))            # temp2 arry grp1
  y1112      = array(apply(y1112Temp2, 3, t), c(nGroup1, 2, numSim)) # y array group 1
  means1112  = apply(y1112, c(3, 2), mean)                           # means mtrx grp1
  covs1112   = array(apply(y1112, 3, cov), c(2, 2, numSim))          # covs array grp1
}

if (nGroup2 > 2) {
  means2122 = mvrnorm(numSim, c(mu21, mu22), popCov2122 / nGroup2)    # means mtrx grp2
  covs2122  = rWishart(numSim, degf2, popCov2122) / degf2             # covs array grp2

} else {
  y2122Temp1 = mvrnorm(numSim*nGroup2, c(mu21, mu22), popCov2122)     # temp1 mtrx grp2
  y2122Temp2 = array(t(y2122Temp1), c(2, nGroup2, numSim))            # temp2 arry grp2
  y2122      = array(apply(y2122Temp2, 3, t), c(nGroup2, 2, numSim)) # y array group 2
  means2122  = apply(y2122, c(3, 2), mean)                           # means mtrx grp2
  covs2122   = array(apply(y2122, 3, cov), c(2, 2, numSim))          # covs array grp2
}

# sample pooled within-group 2x2 covariance matrix for repeated measures
covsPool = (degf1*covs1112 + degf2*covs2122) / degf

# define contrasts for mean comparisons
nAvg = (nGroup1 + nGroup2) / 2
f    = nGroup1 / nAvg
g    = nGroup2 / nAvg

contrast11v12 = c( 1, -1)          # simple effect within group 1
contrast22v21 = c(-1,  1)          # simple effect within group 2
contrast11v21 = c( 1,  0, -1,  0) # simple effect between groups for first measure
contrast22v12 = c( 0, -1,  0,  1) # simple effect between groups for second measure
```

```
# compute sample mean differences
means     = cbind(means1112, means2122)


mean11v21 = means     %*% contrast11v21
mean22v12 = means     %*% contrast22v12
mean11v12 = means1112 %*% contrast11v12
mean22v21 = means2122 %*% contrast22v21

adjTwo = sqrt(nGroup1*nGroup2  / (nGroup1+nGroup2))
adjGr1 = sqrt(nGroup1)
adjGr2 = sqrt(nGroup2)

# contrast for sample covariance matrices
contrastDif = c(1, -1)

# define quadratic forms used to compute variances
quadForm11v12 = function(covX) t(contrast11v12) %*% covX %*% contrast11v12
quadForm22v21 = function(covX) t(contrast22v21) %*% covX %*% contrast22v21
quadFormVar1  = function(covX) covX[1, 1]
quadFormVar2  = function(covX) covX[2, 2]

# compute variances
var11v21 = apply(covsPool, 3, quadFormVar1)
var22v12 = apply(covsPool, 3, quadFormVar2)
var11v12 = apply(covs1112, 3, quadForm11v12)
var22v21 = apply(covs2122, 3, quadForm22v21)

# compute t-statistics
t11v12 = adjGr1 * mean11v12 / sqrt(var11v12) # one-sample t-test within group 1
t22v21 = adjGr2 * mean22v21 / sqrt(var22v21) # one-sample t-test within group 2

t11v21 = adjTwo * mean11v21 / sqrt(var11v21) # two-sample t-test between groups
                                             # at level 1 of within-group factor

t22v12 = adjTwo * mean22v12 / sqrt(var22v12) # two-sample t-test between groups
                                             # at level 2 of within-group factor

# absolute value of t-statistics
absT11v12 = abs(t11v12)
absT11v21 = abs(t11v21)
absT22v12 = abs(t22v12)
absT22v21 = abs(t22v21)

# raw p-values for simple-effect t-statistics
p11v12 = 2*pt(-absT11v12, degf1)
p11v21 = 2*pt(-absT11v21, degf )
p22v12 = 2*pt(-absT22v12, degf )
p22v21 = 2*pt(-absT22v21, degf2)

# matrix of p-values (each row is for a given simulation)
pMat = cbind(p11v12, p11v21, p22v12, p22v21)

# null hypothesis statuses
null11v12  =  mu11 == mu12
null11v21  =  mu11 == mu21
null22v12  =  mu12 == mu22
null22v21  =  mu21 == mu22

# minimum raw p-value for a true null hypothesis
minPTrueNull = pmin(p11v12 + !null11v12, p11v21 + !null11v21,
                    p22v12 + !null22v12, p22v21 + !null22v21)
```

```
# minimum Hommel-adjusted simple-effect p-value for a true null hypothesis
hommelAdjust     = function(p) p.adjust(p, "hom")

minWhereTrueNull = function(p) min(p[c(null11v12, null11v21,
                                       null22v12, null22v21)==TRUE])

# matrix of Hommel-adjusted p-values (each column is for a given simulation)
pHom             = apply(pMat, 1, hommelAdjust    )

# minimum Hommel-adjusted true-null p-value for each simulation
minPHomTrueNull = apply(pHom, 2, minWhereTrueNull)

################################################################################

# BINARY SEARCH FOR OPTIMAL UNIFORM COMPARISONWISE ALPHA LEVEL (aComp)
# FOR SIMPLE-EFFECT TESTS

targetNumSimsError = round(a*numSim) # target number of sims producing >=1 error
aComp              = a/4             # initialize comparisonwise alpha level
stepSize           = aComp/2         # initialize step size to nudge aComp up & down

aCompFound          = FALSE # initialize flag indicating whether optimal cpaComp found
latestStepDirection = 0     # initialize latest step direction (0=down, 1=up)

while (aCompFound==FALSE) {

  # total number of simulations producing at least one Type I error
  numSimsError = sum(minPTrueNull < aComp)

  # nudge aComp up or down as necessary
  if (numSimsError > targetNumSimsError) {          # ERROR RATE IS TOO HIGH
    stepSize = stepSize / (latestStepDirection + 1) # if prev step up, halve step size
    aComp    = max(0, aComp - stepSize)             # nudge down aComp
    latestStepDirection = 0                         # set latest step direction to dwn

  } else if (numSimsError < targetNumSimsError) {   # ERROR RATE IS TOO LOW
    stepSize = stepSize / (2 - latestStepDirection) # if prev step dwn, halve stepsize
    aComp    = aComp + stepSize                     # nudge up aComp
    latestStepDirection = 1                         # set latest step direction to up

  } else {                                          # ERROR RATE IS JUST RIGHT
    aCompFound = TRUE                               # optimal aComp has been found
  }
}

################################################################################

# BINARY SEARCH FOR OPTIMAL NOMINAL FAMILYWISE ALPHA LEVEL (aHom)
# FOR SIMPLE-EFFECT TESTS IN THE ONE-TRACK HOMMEL PROCEDURE

targetNumSimsError  = round(a*numSim) # target number of simulations producing >=1 err
aHom                = a               # initialize nominal familywise Hommel alpha
stepSize            = a/10            # initialize step size to nudge aHom up and down

aHomFound           = FALSE # initialize flag indicating whether optimal aHom found
latestStepDirection = 0     # initialize latest step direction (0=down, 1=up)

while (aHomFound==FALSE) {

  numSimsError = sum(minPHomTrueNull < aHom) # total number of sims producing >=1 err
```

```
  # nudge aHom up or down as necessary
  if (numSimsError > targetNumSimsError) {           # ERROR RATE IS TOO HIGH
    stepSize = stepSize / (latestStepDirection + 1) # if prev step up, halve step size
    aHom     = max(0, aHom - stepSize)              # nudge down aHom
    latestStepDirection = 0                         # set latest step direction to dwn

  } else if (numSimsError < targetNumSimsError) {   # ERROR RATE IS TOO LOW
    stepSize = stepSize / (2 - latestStepDirection) # if prev step dwn, halve stepsize
    aHom     = aHom + stepSize                      # nudge up aHom
    latestStepDirection = 1                         # set latest step direction to up

  } else {                                          # ERROR RATE IS JUST RIGHT
    aHomFound = TRUE                                # optimal aHom has been found
  }
}

totalTime = (proc.time()[3] - startTime) / 60 # total time elapsed in minutes

################################################################################

# REPORT THE PARAMETERS AND RESULTS

totalTime

numSim  # number of simulations
a       # experimentwise alpha level
nGroup1 # number of subjects in group 1
nGroup2 # number of subjects in group 2
aComp   # optimal comparisonwise alpha level for one-track AC method
a/aComp # ratio of experimentwise alpha level to aComp
aHom    # nominal familywise alpha level for one-track Hommel method
```

REFERENCES

Abdi, H., & Williams, L. J. (2010). Contrast analysis. In N. Salkind (Ed.), *Encyclopedia of research design* (pp. 243–251). Thousand Oaks, CA: Sage.

Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Aberson, C. L. (2010). *Applied power analysis for the behavioral sciences*. New York, NY: Routledge.

Ahlbom, A. (1993). *Biostatistics for epidemiologists*. Boca Raton, FL: Lewis Publishers.

Aickin, M. (1999). Other method for adjustment of multiple testing exists. *The BMJ, 318*(7176), 127–128. doi:10.1136/bmj.318.7176.127a

Aickin, M., & Gensler, H. (1996). Adjusting for multiple testing when reporting research results: The Bonferroni vs. Holm methods. *American Journal of Public Health, 86*(5), 726–728. doi:10.2105/AJPH.86.5.726

Albert, J., & Rizzo, M. (2012). *R by example*. New York, NY: Springer.

Allott, K. A., Rapado–Castro, M., Proffitt, T.-M., Bendall, S., Garner, B., Butselaar, F., Markulev, C., Phassouliotis, C., McGorry, P. D., Wood, S. J., Cotton, S. M., & Phillips, L. J. (2015). The impact of neuropsychological functioning and coping style on perceived stress in individuals with first-episode psychosis and healthy controls. *Psychiatry Research, 226*(1), 128–135. doi:10.1016/j.psychres.2014.12.032

American Psychological Association. (2011). *Publication manual of the American Psychological Association* (6th ed.). Washington, D.C.: Author.

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optics, 34*(5), 502–508. doi:10.1111/opo.12131

Aschengrau, A., & Seage, G. R. III. (2014). *Essentials of epidemiology in public health* (3rd ed.). Burlington, MA: Jones & Barlett Learning.

Askari, S., Kirby, R. L., Parker, K., Thompson, K., & O'Neill, J. (2013). Wheelchair propulsion test: Development and measurement properties of a new test for manual wheelchair users. *Archives of Physical Medicine and Rehabilitation, 94*(9), 1690–1698. doi:10.1016/ j.apmr.2013.03.002

Baguley, T. S. (2012). *Serious stats: A guide to advanced statistics*. New York, NY: Palgrave Macmillan.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature, 533*(7604), 452–454. doi:10.1038/533452a

Barnette, J. J., & McLean, J. E. (2005). Type I error of four pairwise mean comparison procedures conducted as protected and unprotected tests. *Journal of Modern Applied Statistical Methods, 4*(2), 446–459. doi:10.22237/jmasm/1130803740

Bechofer, R. E., & Dunnett, C. W. (1982). Multiple comparisons for orthogonal contrasts: Examples and tables. *Technometrics, 24*(3), 213–222. doi:10.1080/00401706.1982.10487761

Bekafigo, M. A., Stepanova, E. V., Eiler, B. A., Noguchi, K., & Ramsey, K. L. (2019). The effect of group polarization on opposition to Donald Trump. *Political Psychology*. doi:10.1111/pops.12584

Bender, R., & Lange, S. (1998). What's wrong with arguments against multiplicity adjustments [letter to the editor]. *BMJ*. Retrieved June 30, 2019, from http://www.bmj.com/rapid-response/2011/10/27/whats-wrong-arguments-against-multiplicity-adjustments

Bender, R. & Lange, S. (2001). Adjusting for multiple testing—when and how? *Journal of Clinical Epidemiology, 54*(4), 343–349. doi:10.1016/S0895-4356(00)00314-0

Benjamini, Y. (2010). Simultaneous and selective inference: Current successes and future challenges. *Biometrical Journal, 52*(6), 708–721. doi:1002/bimj.200900299

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 (Methodology), 57*(1), 289–300. doi:10.2307/2346101

Benjamini, Y., & Yekutieli, D. (2005). False discovery rate: Adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association, 100*(469), 71–81. doi:10.1198/016214504000001907

Berk, M., Daglas, R., Dandash, O., Yücel, M., Henry, L., Hallam, K., Macneil, C., Hasty, M., Pantelis, C., Murphy, B. P., Kader, L., Damodaran, S., Wong, M. T. H., Conus, P., Ratheesh, A., McGorry, P. D., & Cotton, S. M. (2017). Quetiapine v. lithium in the maintenance phase following a first episode of mania: Randomised controlled trial. *The British Journal of Psychiatry, 210*(6), 413–421. doi:10.1192/bjp.bp.116.186833

Berk, M., Dean, O. M., Cotton, S. M., Jeavons, S., Tanious, M., Kohlmann, K., Robbins, J., Cobb, H., Ng, F., Dodd, S., Bush, A. I., & Malhi, G. S. (2014). The efficacy of adjunctive N-acetylcysteine in major depressive disorder: A double-blind, randomized, placebo-controlled trial. *The Journal of Clinical Psychiatry, 75*(6), 628–636. doi:10.4088/JCP.13m08454

Berry, D. (2012). Multiplicities in cancer research: Ubiquitous and necessary evils. *JNCI: Journal of the National Cancer Institute, 104*(15), 1125–1133. doi:10.1093/jnci/djs301

Bethea, R. M., Duran, B. S., & Boullion, T. L. (1995). *Statistical methods for engineers and scientists* (3rd ed.). New York, NY: Marcel Dekker.

Bibby, P. (2012). Simple main effects. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 1376–1380). Thousand Oaks, CA: Sage.

Bird, K. D. (1975). Simultaneous contrast testing procedures for multivariate experiments. *Multivariate Behavioral Research, 10*(3), 343–351. doi:10.1207/s15327906mbr1003_7

Bird, K. D., & Hadzi–Pavlovic, D. (2014). Controlling the maximum familywise Type I error rate in analyses of multivariate experiments. *Psychological Methods, 19*(2), 265–280. doi:10.1037/a0033806

Blakesley, R. E., Mazumdar, S., Dew, M. A., 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*(2), 255–264. doi:10.1037/a0012850

Bobko, P. (1986). A solution to some dilemmas when testing hypotheses about ordinal interactions. *Journal of Applied Psychology, 71*(2), 323–326. doi:10.1037/0021-9010.71.2.323

Boulbes, D. R., Costello, T. J., Baggerly, K. A., Fan, F., Wang, R., Bhattacharya, R., Ye, X., & Ellis, L. M. (2018, April 11). A survey on data reproducibility and the effect of publication process on the ethical reporting of laboratory research. *Clinical Cancer Research*. doi:10.1158/1078-0432.CCR-18-0227

Bradley, A. J., Anderson, K. N., Gallagher, P., & McAllister-Williams, R. H. (2019). The association between sleep and cognitive abnormalities in bipolar disorder. *Psychological Medicine*, 1–8. doi:10.1017/S0033291718004038

Bray, J. H., & Maxwell, S. E. (1982). Analyzing and interpreting significant MANOVAs. *Review of Educational Research, 52*(3), 340–367. doi:10.2307/1170422

Bretz, F., & Westfall, P. H. (2014). Multiplicity and replicability: Two sides of the same coin. *Pharmaceutical Statistics, 13*(6), 343–344. doi:10.1002/pst.1648

Brinkman, T. M., Lown, E. A., Li, C., Olsson, I. T., Marchak, J. G., Stuber, M. L., Vuotto, S., Srivastava, D., Nathan, P. C., Leisenring, W. M., Armstrong, G. T., Robison, L. L., & Krull, K. R. (2019). Alcohol consumption behaviors and neurocognitive dysfunction and emotional distress in adult survivors of childhood cancer: A report from the Childhood Cancer Survivors Study. *Addiction, 114*(2), 226–235. doi:10.1111/add.14439

Brown, S. R. (1990). *Experimental design and analysis*. Newbury Park, CA: Sage.

Byrnes, H. F., Miller, B. A., Grube, J. W., Bourdeau, B., Buller, D. B., Wang-Schweig, M., & Woodall, W. G. (2019). Prevention of alcohol use in older teens: A randomized trial of an online family prevention program. *Psychology of Addictive Behaviors, 33*(1), 1–14. doi:10.1037/adb0000442

Cachelin, F. M., Shea, M., Phimphasone, P., Wilson, G. T., Thompson, D. R., & Striegel, R. H. (2014). Culturally adapted cognitive behavioral guided self-help for binge eating: A feasibility study with Mexican Americans. *Cultural Diversity and Ethnic Minority Psychology, 20*(3), 449–457. doi:10.1037/a0035345

Cameron, D. H., Summerfeldt, L. J., Rowa, K., McKinnon, M. C., Rector, N. A., Richter, M. A., Ornstein, T. J., & McCabe, R. E. (2019). Differences in neuropsychological performance between incompleteness- and harm avoidance-related core dimensions in obsessive-compulsive disorder. *Journal of Obsessive Compulsive and Related Disorders, 22*. doi:10.1016/j.jocrd.2019.100448

Cardinal, R. N., & Aitken, M. R. (2006). *ANOVA for the behavioural sciences researcher*. Mahwah, NJ: Lawrence Erlbaum Associates.

Carral–Fernández, L., González–Blanch, C., Goddard, E., González–Gómez, J., Benito–González, P., & Bustamante–Cruz, E. (2016). Planning abilities in patients with anorexia nervosa compared with healthy controls. *The Clinical Neuropsychologist, 30*(2), 228–242. doi:10.1080/13854046.2016.1147603

Cassidy, S. A., Dimova, R., Giguère, B., Spence, J. R., & Stanley, D. J. (2019). Failing grade: 89% of introduction-to-psychology textbooks that define or explain statistical significance do so incorrectly. *Advances in Methods and Practices in Psychological Science*. doi:10.1177/2515245919858072

Christensen, L. B. (2007). *Experimental methodology* (10th ed.). Boston, MA: Pearson.

Clifford, H. D., Hayden, C. M., Khoo, S., Zhang, G., Le Souëf, P. N., & Richmond, P. (2012). CD46 measles virus receptor polymorphisms influence receptor protein expression and primary measles vaccine responses in naive Australian children. *Clinical and Vaccine Immunology, 19*(5), 704–710. doi:10.1128/CVI.05652-11

Cohen, B. H. (2013). *Explaining psychological statistics* (4th ed.) Hoboken, NJ: John Wiley & Sons.

Cole, D. A., Maxwell, S. E., Arvey, R., & Salas, E. (1994). How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables. *Psychological Bulletin*, *115*(3), 465–474. doi:10.1037/0033-2909.115.3.465

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Cotton, S. M., Gleeson, J. F. M., Alvarez–Jimenez, M., & McGorry, P. D. (2010). Quality of life in patients who have remitted from their first episode of psychosis. *Schizophrenia Research, 121*(1–3), 259–265. doi:10.1016/j.schres.2010.05.027

Cotton, S. M., Lambert, M., Berk, M., Schimmelmann, B. G., Butselaar, F. J., McGorry, P. D., & Conus, P. (2013). Gender differences in first episode psychotic mania. *BMC Psychiatry, 13*(82). doi:10.1186/1471-244X-13-82

Covey, T. J., Shucard, J. L., Shucard, D W. (2019). Working memory training and perceptual discrimination training impact overlapping and distinct neurocognitive processes: Evidence from event-related potentials and transfer of training gains. *Cognition, 182*, 50–72. doi:10.1016/j.cognition.2018.08.012

Cramer, A. O. J., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P. P. P., Waldorp, L. J., & Wagenmakers, E.-J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review, 23*, 640–647. doi:10.3758/s13423-015-0913-5

Crawley, M. J. (2013). *The R book* (2nd ed.). Chichester, UK: John Wiley & Sons.

CREDO. (n.d.). CREDO response to critique for multiple comparisons adjustment. Retrieved June 30, 2019, from http://credo.stanford.edu/pdfs/CREDOmethodsresponse.pdf

Cribbie, R. A., & Keselman, H. J. (2003). The effects of nonnormality on parametric, nonparametric, and model comparison approaches to pairwise comparisons. *Educational and Psychological Measurement, 63*(4), 615–635. doi:10.1177/0013164403251283

Dar, R., Serlin, R. C., & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting and Clinical Psychology*, *62*(1), 75–82. doi:10.1037/0022-006X.62.1.75

Day, R. W., & Quinn, G. P. (1989). Comparisons of treatments after an analysis of variance in ecology. *Ecological Monographs, 59*(4), 433–463. doi:10.2307/1943075

Day, M. A., & Thorn, B. E. (2017). Mindfulness-based cognitive therapy for headache pain: An evaluation of the long-term maintenance of effects. *Complementary Therapies in Medicine, 33*, 94–98. doi:10.1016/j.ctim.2017.06.009

De Pablo-Fernandez, E., Tur, C., Revesz, T., Lees, A. J., Holton, J. L., & Warner, T. T. (2017). Association of autonomic dysfunction with disease progression and survival in Parkinson disease. *JAMA Neurology, 74*(8), 970–976. doi:10.1001/jamaneurol.2017.1125

Devore, J. L. (1987). *Probability and statistics for engineering and the sciences* (2nd ed.). Monterey, CA: Brooks/Cole.

Dickhaus, T. (2014). *Simultaneous statistical inference*. Berlin, Germany: Springer.

Dmitrienko, A., Bretz, F., Westfall, P. H., Troendle, J., Wiens, B. L., Tamhane, A. C., & Hsu, J. C. (2010a). Multiple testing methodology. In A. Dmitrienko, A. C. Tamhane, & F. Bretz (Eds.), *Multiple testing problems in pharmaceutical statisics* (pp. 35–98). Boca Raton, FL: Chapman & Hall.

Dmitrienko, A., Tamhane, A. C., & Bretz, F. (Eds.). (2010b). *Multiple testing problems in pharmaceutical statistics*. Boca Raton, FL: Chapman & Hall.

Doncaster, C. P., & Davey, A. J. H. (2007). *Analysis of variance and covariance: How to choose and construct models for the life sciences*. Cambridge, U.K.: Cambridge University Press.

Dondaine, T., Philippot, P., Batail, J.-M., Le Jeune, F., Sauleau, P., Drapier, S., Vérin, M., Millet, B., Drapier, D., & Robert, G. (2019). Apathy alters emotional arousal in chronic schizophrenia. *Journal of Psychiartry & Neuroscience, 44*(1), 54–61.

Dovgan, K., Mazurek, M. O. (2019). Impact of multiple co-occuring emotional and behavioural conditions on children with autism and their families. *Journal of Applied Research in Intellectual Disabilities, 32*(4), 967–980. doi:10.1111/jar.12590

Duffy, A., Dawson, D. L., Moghaddam, N. G., & das Nair, R. (2019). Do thinking styles play a role in whether people pathologise their pornography use? *Sexual and Relationship Therapy, 34*(1), 87–108. doi:10.1080/14681994.2017.1412417

Duncan, D. B. (1951). A significance test for differences between ranked treatments in an analysis of variance. *Virginia Journal of Science, 2*, 171–189.

Duncan, D. B. (1955). Multiple range and multiple *F* tests. *Biometrics, 11*(1), 1–42. doi:10.2307/3001478

Dunn, O. J. (1958). Estimation of the means for dependent variables. *Annals of Mathematical Statistics, 29*(4), 1095–1111.

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

Dunnett. C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association, 50*(272), 1096–1121. doi:10.1080/01621459.1955.10501294

Edwards, D., & Berry, J. T. (1987). The efficiency of simulation-based multiple comparisons. *Biometrics, 43*(4), 913–928.

Eichstaedt, K. E., Kovatch, K., & Maroof, D. A. (2013). A less conservative method to adjust for familywise error rate in neuropsychological research: The Holm's sequential Bonferroni procedure. *NeuroRehabilitation, 32*(3), 693–696. doi:10.3323/NRE-130893

Enders, C. K. (2003). Performing multivariate group comparisons following a statistically significant MANOVA. *Measurement and Evaluation in Counseling and Development*, *36*(1), 40–56.

European Agency for the Evaluation of Medicinal Products. (2002). *Points to consider on multiplicity issues in clinical trials*. Retrieved June 30, 2019, from http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003640.pdf

Falconer, C. J., Lobmaier, J. S., Christoforou, M., Kamboj, S. K., King, J. A., Gilbert, P., & Brewin, C. R. (2019). Compassionate faces: Evidence for distinctive facial expressions associated with specific prosocial motivations. *PLoS ONE, 14*(1), e0210283. doi:10.1371/journal.pone.0210283

Feise, R. J. (2002). Do multiple outcome measures require p-value adjustment? *BMC Medical Research Methodology, 2*(8). doi:10.1186/1471-2288-2-8

Fekkes, M., Pijpers, F. I. M., Fredriks, A. M., Vogels, T., & Verloove-Vanhorick, S. P. (2006). Do bullied children get ill, or do ill children get bullied? A prospective cohort study on the relationship between bullying and health-related symptoms. *Pediatrics, 117*(5), 1568–1574. doi:10.1542/peds.2005-0187

Félix, V. B., & Menezes, A. F. B. (2018). Comparisons of ten corrections methods for *t*-test in multiple comparisons via Monte Carlo study. *Electronic Journal of Applied Statistical Analysis, 11*(1), 74–91. doi:10.1285/i20705948v11n1p74

Fenesi, B., Heisz, J. J., Savage, P. I., Shore, D. I., & Kim, J. A. (2014). Combining best-practice and experimental approaches: Redundancy, images, and misperceptions in multimedia learning. *The Journal of Experimental Education, 82*(2), 253–263. doi:10.1080/00220973.2012.745472

Finley, J. R., Benjamin, A. S., Hays, M. J., Bjork, R. A., & Kornell, N. (2011). Benefits of accumulating versus diminishing cues in recall. *Journal of Memory and Language, 64*(4), 289–298. doi:10.1016/j.jml.2011.01.006

Finner, H., & Roters, M. (2001). On the false discovery rate and expected Type I errors. *Biometrical Journal, 43*(8), 985–1005. doi:10.1002/1521-4036(200112)43:8>985::AID-BIMJ985>3.0CO;2-4

Fish, J., Evans, J. J., Nimmo, M., Martin, E., Kersel, D., Bateman, A., Wilson, B. A., & Manly, T. (2007). Rehabilitation of executive dysfunction following brain injury: "Content-free" cueing improves everyday prospective memory performance. *Neuropsychologia, 45*(6), 1318–1330. doi:10.1016/j.neuropsychologia.2006.09.015

Fisher, R. A. (1935). *The design of experiments*. Edinburgh, U.K.: Oliver & Boyd.

Forstmeier, W., Wagenmakers, E.-J., & Parker, T. H. (2016). Detecting and avoiding likely false-positive findings—a practical guide. *Biological Reviews, 92*(4), 1941–1968. doi:10.1111/brv.12315

Frane, A. V. (2015a). Are per-family Type I error rates relevant in social and behavioral science? *Journal of Modern Applied Statistical Methods, 14*(1), 12–23. doi:10.22237/jmasm/1430453040

Frane, A. V. (2015b). Planned hypothesis tests are not necessarily exempt from multiplicity adjustment. *Journal of Research Practice, 11*(1), P2. Retrieved June 30, 2019, from https://files.eric.ed.gov/fulltext/EJ1083896.pdf

Frane, A. V. (2015c). Power and type I error control for univariate comparisons in multivariate two-group designs. *Multivariate Behavioral Research, 50*(2), 233–247. doi:10.1080/00273171.2014.968836

Frane, A. V. (2016). False discovery rate control is not always a replacement for Bonferroni adjustment. *Journal of Clinical Epidemiology, 69*(1), 263. doi:10.1016/j.jclinepi.2015.03.025

Frenette, L. C., Tinawi, S., Correa, J. A., Alturki, A. Y., LeBlanc, J., Feyz, M., & de Guise, E. (2019). Early detection of cognitive impairments with the Montreal Cognitive Assessment in patients with uncomplicated and complicated mild traumatic brain injury. *Brain Injury, 33*(2), 189–197. doi:10.1080/02699052.2018.1542506

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

Games, P. A., & Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal n's and/or variances: A Monte Carlo study. *Journal of Educational Statistics, 1*(2), 113–125. doi:10.3102/10769986001002113

Girden, E. R. (1992). *ANOVA: Repeated measures*. Newbury Park, CA: Sage.

Glaus, J., Vandeleur, C. L., von Känel, R., Lasserre, A. M., Strippoli, M. F., Gholam–Rezaee, M., Castelao, E., Marques–Vidal, P., Bovet, P., Merikangas, K., Mooser, V., Waeber, G., Vollenweider, P., Aubry, J.-M., & Preisig, M. (2014). Associations between mood, anxiety or substance use disorders and inflammatory markers after adjustment for multiple covariates in a population-based study. *Journal of Psychiatric Research, 58*, 36–

45. doi:10.1016/j.jpsychires.2014.07.012

Glickman, M. E., Rao, S. R., & Schultz, M. R. (2014). False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *Journal of Clinical Epidemiology, 67*(8), 850–857. doi:10.1016/j.jclinepi.2014.03.012

Goeman, J. J., & Solari, A. (2014). Multiple hypothesis testing in genomics. *Statistics in Medicine, 33*(11), 1946–1978. doi:10.1002/sim.6082

Gonzalez, R. (2012). *Data analysis for experimental design*. New York, NY: Guilford.

González–Blanch, C., Gleeson, J. F., Cotton, S. M., Crisp, K., McGorry, P. D., & Alvarez–Jimenez, M. (2015). Longitudinal relationships between expressed emotion and cannabis misuse in young people with first-episode psychosis. *European Psychiatry, 30*(1), 20–25. doi:10.1016/j.europsy.2014.07.002

Goodwin, C. J. (2010). *Research in psychology: Methods and design* (6th ed.). Hoboken, NJ: John Wiley & Sons.

Gordon, A., Glazko, G., & Yakovlev, A. (2007). Control of the mean number of false discoveries, Bonferroni and stability of multiple testing. *The Annals of Applied Statistics*, *1*(1), 179–190. doi:10.1214/07AOAS102

Grayson, D. (2004). Some myths and legends in quantitative psychology. *Understanding Statistics*, *3*(1), 101–134. doi:10.1207/s15328031us0302_3

Grice, J. W., & Iwaski, M. (2007). A truly multivariate approach to MANOVA. *Applied Multivariate Research, 12*(3), 199–226.

Groarke, J. M., & Hogan, M. J. (2019). Listening to self-chosen music regulates induced negative affect for both younger and older adults. *PLoS ONE, 14*(6), e0218017. doi:10.1371/journal.pone.0218017

Guilbaud, O. (2008). Simultaneous confidence regions corresponding to Holm's stepdown procedure and other closed-testing procedures. *Biometrical Journal, 50*(5), 678–692. doi:10.1002/bimj.200710449

Guilbaud, O. (2012). Simultaneous confidence regions for closed tests, including Holm-, Hochberg-, and Hommel-related procedures. *Biometrical Journal, 54*(3), 317–342. doi:10.1002/bimj.201100123

Ha, R. R., & Ha, J. C. (2012). *Integrative statistics for the social and behavioral sciences*. Thousand Oaks, CA: Sage.

Haase, R. F., & Ellis, M. V. (1987). Multivariate analysis of variance. *Journal of Counseling Psychology, 34*(4), 404–413. doi:10.1037/0022-0167.34.4.404

Hackford, J., Mackey, A., & Broadbent, E. (2019). The effects of walking posture on affective and psychological states during stress. *Journal of Behavior Therapy and Experimental Psychiatry, 62*, 80–87. doi:10.1016/j.jbtep.2018.09.004

Hair, J. F., Black, B., Babin, B., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis* (6th ed.). Upper Saddle River, NJ: Pearson.

Hakstian, A. R., Roed, J. C., & Lind, J. C. (1979). Two-sample T^2 procedure and the assumption of homogeneous covariance matrices. *Psychological Bulletin, 86*(6), 1255–1263. doi:10.1037/0033-2909.86.6.1255

Hancock, G. R., & Klockars, A. J. (1996). The quest for α: Developments in multiple comparison procedures in the quarter century since Games (1971). *Review of Educational Research, 66*(3), 269–306. doi:10.3102/00346543066003269

Harkness, K. L., & Luther, J. (2001). Clinical risk factors for the generation of life events in major depression. *Journal of Abnormal Psychology, 110*(4), 564–572. doi:10.1037/0021-843X.110.4.564

Harris, R. J. (2001). *A primer of multivariate statistics* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Hays, W. L. (1988). *Statistics* (4th ed.). Fort Worth, TX: Holt, Rinehart and Winston.

Hayter, A. J. (1986). The maximum familywise error rate of Fisher's least significant difference test. *Journal of the American Statistical Association, 81*(396), 1000–1004. doi:10.1080/01621459.1986.10478364

Heiberger, R. M., & Holland, B. (2004). *Statistical analysis and data display: An intermediate course with examples in S-Plus, R, and SAS*. New York, NY: Springer.

Helweg–Larsen, M., & Nielsen, G. A. (2009). Smoking cross-culturally: Risk perceptions among young adults in Denmark and the United States. *Psychology and Health, 24*(1), 81–93. doi:10.1080/08870440801932656

Hinton, P. R. (2004). *Statistics explained* (2nd ed.). New York, NY: Routledge.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*(4), 800–802. doi:10.1093/biomet/75.4.800

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York, NY: John Wiley & Sons.

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

Holmes, A. L., Joyce, K., Xie, H., Falank, C., Hinz, J. M., & Wise, J. P., Sr. (2014). The impact of homologous recombination repair deficiency on depleted uranium clastogenicity in Chinese hamster ovary cells: XRCC3 protects cells from chromosome aberrations, but increases chromosome fragmentation. *Mutation Research: Fundamental and Molecular Mechanisms of Mutagenesis, 762*, 1–9. doi:10.1016/j.mrfmmm.2014.02.001

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika, 75*(2), 383–386. doi:10.1093/biomet/75.2.383

Howell, D. C. (2013). *Statistical methods for psychology* (8th ed.). Belmont, CA: Wadsworth.

Hsu, J. C. (1996). *Multiple comparisons*. Boca Raton, FL: Chapman & Hall.

Hsu, J. C., & Nelson, B. (1998). Multiple comparisons in the general linear model. *Journal of Computational and Graphical Statistics, 7*(1), 23–41. doi:10.1080/10618600.1998.10474759

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

Huberty, C. J., & Petoskey, M. S. (2000). Multivariate analysis of variance and covariance. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 183–208). San Diego, CA: Academic Press.

Huisingh, C., & McGwin, G., Jr. (2012). An analysis of the use of multiple comparison corrections in ophthalmology research [letter to the editor]. *Investigative Ophthalmology & Visual Science, 53*(8), 4777. doi:10.1167/iovs.12-10336

Hummel, T. J., & Sligo, J. R. (1971). Empirical comparison of univariate and multivariate analysis of variance procedures. *Psychological Bulletin, 76*(1), 49–57. doi:10.1037/h0031323

Hurlbert, S. H., & Lombardi, C. M. (2012). Lopsided reasoning on lopsided tests and multiple comparisons. *Australian & New Zealand Journal of Statistics, 54*(1), 23–42. doi:10.1111/j.1467-842X.2012.00652.x

Iacobucci, D. (2001). Analysis of variance: I.A. Can I test for simple effects in the presence of an insignificant interaction? *Journal of Consumer Psychology, 10*(1–2), 5–9. doi:10.1207/S15327663JCP1001&2_03

Jacka, F. N., O'Neil, A., Opie, R., Itsiopoulos, C., Cotton, S., Mohebbi, M., Castle, D., Dash, S., Mihalopoulos, C., Chatterton, M. L., Brazionis, L., Dean, O. M., Hodge, A. M., & Berk, M. (2017). A randomized controlled trial of dietary improvement for adults with major depression (the 'SMILES' trial). *BMC Medicine, 15*(23). doi:10.1186/s12916-017-0791-y

Jamieson, K. H. (2018). Crisis or self-correction: Rethinking media narratives about the well-being of science. *Proceedings of the National Academy of Sciences of the United States of America*. doi:10.1073/pnas.1708276114

Jenkins, T. M., Toosy, A. T., Ciccarelli, O., Miszkiel, K. A., Wheeler-Kingshott, C. A., Henderson, A. P., Kallis, C., Mancini, L., Plant, G. T., Miller, D. H., & Thompson, A. J. (2009). Neuroplasticity predicts outcome of optic neuritis independent of tissue damage. *Annals of Neurology, 67*(1), 99–113. doi:10.1002/ana.21823

Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Jung, Y.-H., Jang, J. H., Lee, D., Choi, Y., Choi, S.-H., Kang, D.-H. (2019). Relationships between catecholamine levels and stress or intelligence. *Neurochemical Research, 44*(5), 1192–1200. doi:10.1007/s11064-019-02762-z

Kaseweter, K., Rose, K., Bednarik, S., & Woodworth, M. (2019). More than meets the eye: The role of psychopathic traits in attention to distress. *Psychology, Crime, & Law*. doi:0.1080/1068316X.2019.1634198

Kawai, V. K., Avalos, I., Oeser, A., Oates, J. A., Milne, G. L., Solus, J., Chung, C. P., & Stein, C. M. (2014). Suboptimal inhibition of platelet cyclooxygenase 1 by aspirin in systemic lupus erythematosus: Association with metabolic syndrome. *Arthritis Care and Research, 66*(2), 285–292. doi:10.1002/acr.22169

Kellow, J. T. (2000). Misuse of multivariate analysis of variance in behavioral research: The fallacy of the "protected" *F* test. *Perceptual and Motor Skills, 90*(3), 917–926. doi:10.2466/PMS.90.3.917–926

Keppel, G., & Zedeck, S. (1989). *Data analysis for research designs*. New York, NY: W. H. Freeman & Company.

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*(1), 58–69. doi:10.1037/1082-989X.4.1.58

Keselman, H, J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A, Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*(3), 350–386. doi:10.3102/00346543068003350

Keuls, M. (1952). The use of Studentized range in connection with an analysis of variance. *Euphytica, 1*(2), 112–122.

Kim, W. B., Alavi, A., Walsh, S., Kim, S., & Pope, E. (2015). Epidermolysis bullosa pruriginosa: A systematic review exploring genotype–phenotype correlation. *American Journal of Clinical Dermatology, 16*(2), 81–87. doi:10.1007/s40257-015-0119-7

Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences* (4th ed.). Thousand Oaks, CA: Sage.

Klockars, A. J., & Hancock, G. R. (1994). Per-experiment error rates: The hidden costs of several multiple comparison procedures. *Educational and Psychological Measurement*, *54*(2), 292–298. doi:10.1177/0013164494054002004

Klockars, A. J., Hancock, G. R., & McAweeney, M. J. (1995). Power of unweighted and weighted versions of simultaneous and sequential multiple-comparison procedures. *Psychological Bulletin, 118*(2), 300–307. doi:10.1037/0033-2909.118.2.300

Koegi, C. J. (2019). A short and long-term evaluation of a substance abuse program for incarcerated men. *Journal of Offender Rehabilitation, 58*(4), 281–304. doi:10.1080/10509674.2019.1596192

Kramer, C. Y. (1956). Extensions of multiple range tests to group means with unequal number of replications. *Biometrics, 12*(3), 307–310. doi:10.2307/3001469

Krane–Gartiser, K., Henriksen, T. E. G., Morken, G., Vaaler, A., & Fasmer, O. B. (2014). Actigraphic assessment of motor activity in acutely admitted inpatients with bipolar disorder. *PLoS ONE, 9*(2), e89574. doi:10.1371/journal.pone.0089574

Kromrey, J. D., & Dickinson, W. B. (1995). The use of an overall *F* test to control Type I error rates in factorial analyses of variance: Limitations and better strategies. *The Journal of Applied Behavioral Science, 31*(1), 51–64. doi:10.1177/0021886395311006

Lange, B. P., von Andrian–Werburg, M. T. P., Adler, D. C., & Zaretsky, E. (2019). The name is the game: Nicknames as predictors of personality and mating strategy in online dating. *Frontiers in Communication, 4*, Article 3. doi:10.3389/fcomm.2019.00003

Larrabee, M. J. (1982). Reexamination of a plea for multivariate analyses. *Journal of Counseling Psychology, 29*(2), 180–188. doi:10.1037/0022-0167.29.2.180

Lau, M., Lin, H., & Flores, G. (2012). Racial/ethnic disparities in health and health care among U.S. adolescents. *Health Services Research, 47*(5), 2031–2059. doi:10.1111/j.1475-6773.2012.01394.x

Leary, M. R., & Altmaier, E. M. (1980). Type I error in counseling research: A plea for multivariate analyses. *Journal of Counseling Psychology, 27*(6), 611–615. doi:10.1037/0022-0167.27.6.611

Levin, B. (1996). On the Holm, Simes, and Hochberg multiple test procedures. *American Journal of Public Health, 86*(5), 628–629. doi:10.2105/AJPH.86.5.628

Levin, J. R., Serlin, R. C., & Seaman, M. A. (1994). A controlled, powerful multiple-comparison strategy for several situations. *Psychological Bulletin, 115*(1), 153–159. doi:10.1037/0033-2909.115.1.153

Li, D. (2008). A two-step rejection procedure for testing multiple hypotheses. *Journal of Statistical Planning and Inference, 138*(6), 1521–1527. doi:10.1016/j.jspi.2007.04.032

Li, L. Y., Karcher, N. R., Kerns, J. G., Fung, C. K., & Martin, E. A. (2019). The subjective–objective deficit paradox in schizotypy extends to emotion regulation and awareness. *Journal of Psychiatric Research, 111*, 160–168. doi:10.1016/j.jpsychires.2019.01.026

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology, 54*(1), 146–157. doi:10.1111/psyp.12639

Ludbrook, J. (1991). On making multiple comparisons in clinical and experimental pharmacology and physiology. *Clinical and Experimental Pharmacology and Physiology, 18*, 379–392. doi:10.1111//j.1440-1681.1991.tb01468.x

Ludbrook, J. (1998). Multiple comparison procedures updated. *Clinical and Experimental Pharmacology and Physiology, 25*, 1032–1037. doi:10.1111/j.1440-1681.1998.tb02179.x

MacDonald, B., & Barry, R. J. (2014). Trial effects in single-trial ERP components and autonomic responses at very long ISIs. *International Journal of Psychophysiology, 92*(3), 99–112. doi:10.1016/j.ijpsycho.2014.03.007

Marion–Veyron, R., Lambert, M., Cotton, S. M., Schimmelmann, B. G., Gravier, B., McGorry, P. D., & Conus, P. (2015). History of offending behavior in first episode psychosis patients: A marker of specific clinical needs and a call for early detection strategies among young offenders. *Schizophrenia Research, 161*(2–3), 163–168. doi:10.1016/j.schres.2014.09.078

McHugh, R. B., & Ellis, D. S. (1957). The "post-mortem" testing of experimental comparisons. *Psychological Bulletin, 52*(5), 425–428. doi:10.1037/h0044530

McKillup, S. (2012). *Statistics explained: An introductory guide for life scientists*. Cambridge, U.K.: Cambridge University Press.

Meijer, R. J., & Goeman, J. J. (2016). Multiple testing of gene sets from gene ontology: Possibilities and pitfalls. *Briefings in Bioinformatics, 17*(5), 808–818. doi:10.1093/bib/bbv091

Mertler, C. A., & Vannatta, R. A. (2010). *Advanced and multivariate statistical methods* (4th ed.). Glendale, CA: Pyrczak.

Meyers, L. S., Gamst, G., & Guarino, A. J. (2006). *Applied multivariate research*. Thousand Oaks, CA: Sage.

Miles, J., & Banyard, P. (2007). *Understanding and using statistics in psychology*. London, U.K.: Sage.

Miller, R. G., Jr. (1966). *Simultaneous statistical inference*. New York, NY: McGraw–Hill.

Miller, R. G., Jr. (1981). *Simultaneous statistical inference* (2nd ed.). New York, NY: Springer–Verlag.

Moran, M. D. (2003). Arguments for rejecting the sequential Bonferroni in ecological studies. *Oikos, 100*, 403–405. doi:10.1034/j.1600-0706.2003.12010.x

Morrison, D. F. (1967). *Multivariate statistical methods*. New York, NY: McGraw–Hill.

Mossaheb, N., Schäfer, M. R., Schlögelhofer, M., Klier, C. M., Cotton, S. M., McGorry, P. D., & Amminger, G. P. (2013). Effect of omega-3 fatty acids for indicated prevention of young patients at risk for psychosis: When do they begin to be effective? *Schizophrenia Research, 148*(1–3), 163–167. doi:10.1016/j.schres.2013.05.027

Nakagawa, S. (2004). A farewell to Bonferroni: The problems of low statistical power and publication bias. *Behavioral Ecology, 15*(6), 1044–1045. doi:10.1093/beheco/arh107

Nakamae, T., Sakai, Y., Abe, Y., Nishida, S., Fukui, K., Yamada, K., Kubota, M., Denys, D., & Narumoto, J. (2014). Altered fronto-striatal fiber topography and connectivity in obsessive-compulsive disorder. *PLoS ONE, 9*(11), e112075. doi:10.1371/journal.pone.0112075

Nam, C. W., & Zellner, R. D. (2011). The relative effects of positive interdependence and group processing on student achievement and attitude in online cooperative learning. *Computers and Education, 56*(3), 680–688. doi:10.1016/j.compedu.2010.10.010

Nestor, P. G., & Schutt, R. K. (2012). *Research methods in psychology: Investigating human behavior*. Thousand Oaks, CA: Sage.

Newman, D. (1939). The distribution of the range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. *Biometrika, 31*(1–2), 20–30. doi:10.2307/2334973

Nieuwenhuis, I. L. C., Folia, V., Forkstam, C., Jensen, O., & Petersson, K. M. (2013). Sleep promotes the extraction of grammatical rules. *PLoS ONE, 8*(6), e65046. doi:10.1371/journal.pone.0065046

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods, 48*(4), 1205–1226. doi:10.3758/s13428-015-0664-2

O'Connor, M., Harris, J. M., McIntosh, A. M., Owens, D. G. C., Lawrie, S. M., & Johnstone, E. C. (2009). Specific cognitive deficits in a group at genetic high risk of schizophrenia. *Psychological Medicine, 39*, 1649–1655. doi:10.1017/S0033291709005303

O'Keefe, D. J. (2003). Colloquy: Should familywise alpha be adjusted? Against familywise alpha level adjustment. *Human Communication Research, 29*(3), 431–447. doi:10.1111/j.1468-2958.2003.tb00846.x

Olejnik, S., & Hess, B. (1997). Top ten reasons why most omnibus ANOVA *F* tests should be abandoned. *Journal of Vocational Educational Research, 22*(4), 219–232.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. doi:10.1126/science.aac4716

Ostendorf, D. M., Lyden, K., Pan, Z., Wyatt, H. R., Hill, J. O., Melanson, E. L., & Catenacci, V. A. (2017). Objectively measured physical activity and sedentary behavior in successful weight loss maintainers. *Obesity, 26*(1), 53–60. doi:10.1002/oby.22052

Ozcan, T., Bacak, S. J., Zozzaro-Smith, P., Li, D., Sagcan, S., Seligman, N., & Glantz, C. J. (2017). Assessing weight gain by the 2009 Institute of Medicine guidelines and perinatal outcomes in twin pregnancy. *Maternal and Child Health Journal, 21*(3), 509–515. doi:10.1007/s10995-016-2134-6

Pagano, R. R. (2013). *Understanding statistics in the behavioral sciences* (10th ed.). Boston, MA: Cengage.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*(6). doi:10.1177/1745691612465253

Pataki, K. W., Metz, A. E., & Pakulski, L. (2014). The effect of thematically related play on engagement in storybook reading in children with hearing loss. *Journal of Early Childhood Literacy, 14*(2), 240–264. doi:10.1177/1468798413480516

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design and analysis: An integrated approach*. New York, NY: Psychology Press.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *The BMJ, 316*(7139), 1236–1238. doi:10.1136/bmj.316.7139.1236

Phillips, A., Fletcher, C., Atkinson, G., Channon, E., Douiri, A., Jaki, T., Maca, J., Morgan, D., Roger, J. H., & Terrill, P. (2013). Multiplicity: Discussion points from the Statisticians in the Pharmaceutical Industry multiplicity expert group. *Pharmaceutical Statistics, 12*(5), 255–259. doi:10.1002/pst.1584

Posch, M., & Futschik, A. (2012). A uniform improvement of Bonferroni-type tests by sequential tests. *Journal of the American Statistical Association, 103*(481), 299–308. doi:10.1198/016214508000000012

Pyra, M., Weber, K., Wilson, T. E., Cohen, J., Murchison, L., Goparaju, L., & Cohen, M. H. (2014). Sexual minority status and violence among HIV-infected and at-risk women. *Journal of General Internal Medicine, 29*(8), 1131–1138. doi:10.1007/s11606-014-2832-y

Racette, L., Boden, C., Kleinhandler, S. L., Girkin, C. A., Liebmann, J. M., Zangwill, L. M., Medeiros, F. A., Bowd, C., Weinreb, R. M., Wilson, M. R., & Sample, P. A. (2005). Differences in visual function and optic nerve structure between healthy eyes of Blacks and Whites. *JAMA Ophthalmology, 123*(11), 1547–1553. doi:10.1001/archopht.123.11.1547

Rajapakse, T., Griffiths, K. M., Christensen, H., & Cotton, S. (2014). A comparison of non-fatal self-poisoning among males and females, in Sri Lanka. *BMC Psychiatry, 14*(221). doi:10.1186/s12888-014-0221-z

Ramsey, P. H. (1978). Power differences between pairwise multiple comparisons. *Journal of the American Statistical Association, 73*(363), 479–485. doi:10.1080/01621459.1978.10480038

Ramsey, P. H. (1982). Empirical power of procedures for comparing two groups on *p* variables. *Journal of Educational Statistics*, *7*(2), 139–156. doi:10.2307/1164962

Ramsey, P. H., Barrera, K., Hachimine–Semprebom, P., & Li, C.-C. (2011). Pairwise comparisons of means under realistic nonnormality, unequal variances, outliers and equal sample sizes. *Journal of Statistical Computation and Simulation, 81*(2), 125–135. doi:10.1080/00949650903219935

Ramsey, P. H., & Ramsey, P. P. (2008). Power of pairwise comparisons in the equal variance and unequal sample size case. *British Journal of Mathematical and Statistical Psychology, 61*(1), 115–131. doi:10.1348/000711006X153051

Ramsey, P. H., & Ramsey, P. P. (2009). Power and Type I errors for pairwise comparisons of means in the unequal variances case. *British Journal of Mathematical and Statistical Psychology, 62*(2), 263–281. doi:10.1348/000711008X291542

Randolph, K. A., & Meyers, L. L. (2013). *Basic statistics in multivariate analysis*. Oxford, U.K.: Oxford University Press.

Ray, N. J., Brittain, J. S., Holand, P., Joundi, R. A., Stein, J. F., Aziz, T. Z., & Jenkinson, N. (2012). The role of the subthalamic nucleus in response inhibition: Evidence from local field potential recordings in the human subthalamic nucleus. *Neuroimage, 60*(1), 271–278. doi:10.1016/j.neuroimage.2011.12.035

R Core Team. (2013). R: A language and environment for statistical computing. https://www.R-project.org/

R Core Team. (2017). R: A language and environment for statistical computing. https://www.R-project.org/

Richter, S. J., & McCann, M. H. (2012). Using the Tukey–Kramer omnibus test in the Hayter–Fisher procedure. *British Journal of Mathematical and Statistical Psychology, 65*(3), 499–510. doi:10.1111/j.2044-8317.2012.02041.x

Roberts, J., Williams, K., Carter, M., Evans, D., Parmenter, T., Silove, N., Clark, T., & Warren, A. (2011). A randomised controlled trial of two early intervention programs for young children with autism: Centre-based with parent program and home-based. *Research in Autism Spectrum Disorders, 5*(4), 1553–1566. doi:10.1016/j.rasd.2011.03.001

Roche, K., & Chainay, H. (2013). Visually guided grasping of common objects: Effects of priming. *Visual Cognition, 21*(8), 1010–1032. doi:10.1080/13506285.2013.851136

Rom, D. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika*, *77*(3), 663–665. doi:10.1093/biomet/77.3.663

Rosnow, R. L., & Rosenthal, R. (1989). Definition and interpretation of interaction effects. *Psychological Bulletin, 105*(1), 143–146. doi:10.1037/0033-2909.105.1.143

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

Rovai, A. P., Baker, J. D., & Ponton, M. K. (2014). *Social science research design and statistics: A practitioner's guide to research methods and IBM SPSS analysis*. Chesapeake, VA: Waterfree Press.

Rubin, M. (2017). Do p values lose their meaning in exploratory analyses? It depends how you define the familywise error rate. *Review of General Psychology, 21*(3), 269–275. doi:10.1037/gpr0000123

Ruxton, G. D., & Beauchamp, G. (2008). Time for some a priori thinking about post hoc testing. *Behavioral Ecology, 19*(3), 690–693. doi:10.1093/beheco/arn020

Rutherford, A. (2011). *ANOVA and ANCOVA: A GLM approach* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Ryan, T. A. (1959). Multiple comparisons in psychological research. *Psychological Bulletin*, *56*(1), 26–47. doi:10.1037/h0042478

Ryan, T. A. (1962). The experiment as the unit for computing rates of error. *Psychological Bulletin*, *59*(4), 301–305. doi:10.1037/h0040562

Ryan, T. A. (1995, December 1). 'Post hoc' tests [posted to STAT-L discussion group]. Retrieved June 30, 2019, from http://groups.google.com/forum/#!activity/sci.stat.consult/JznJOuHfr7QJ/sci.stat.consult/fJmYD9TJQ6A/AfMk-8gvJacJ

Samuel–Cahn, E. (1996). Is the Simes improved Bonferroni procedure conservative? *Biometrika*, *83*(4), 928–933. doi:10.1093/biomet/83.4.928

Sankoh, A. J., Huque, M. F., & Dubey, S. D. (1997). Some comments on frequently used multiple endpoint adjustment methods in clinical trials. *Statistics in Medicine, 16*, 2529–2542. doi:10.1002/(SICI)1097-0258(19971130)16:223.0.CO;2-J

Sauder, D. C., & DeMars, C. E. (2019). An updated recommendation for multiple comparisons. *Advances in Methods and Practices in Psychological Science, 2*(1), 26–44. doi:10.1177/2515245918808784

Saville, D. J. (1990). Multiple comparison procedures: The practical solution. *The American Statistician, 44*(2),174–180. doi:10.2307/2684163

Savitz, D. A. (2003). *Interpreting epidemiologic evidence: Strategies for study design and analysis*. Oxford, U.K.: Oxford University Press.

Savitz, D. A., & Olshan, A. F. (1995). Multiple comparisons and related issues in the interpretation of epidemiologic data. *American Journal of Epidemiology, 142*(9), 904–908. doi:10.1093/oxfordjournals.aje.a117737

Scheff, S. W. (2016). *Fundamental statistical principles for the neurobiologist: A survival guide*. London, U.K.: Academic Press.

Schmitt, J. A. J., Jorissen, B. L., Sobczak, S., van Boxtel, M. P. J., Hogervorst, E., Deutz, N. E. P., & Riedel, W. J. (2000). Tryptophan depletion impairs memory consolidation but improves focussed attention in healthy young volunteers. *Journal of Psychopharmacology, 14*(1), 21–29. doi:10.1177/026988110001400102

Seaman, M. A., Levin, J. R., & Serlin, R. C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures. *Psychological Bulletin*, *110*(3), 577–586. doi:10.1037/0033-2909.110.3.577

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association, 81*(395), 826–831. doi:10.1080/01621459.1986.10478341

Shaffer, J. P. (1991). Probability of directional errors with disordinal (qualitative) interaction. *Psychometrika, 56*(1), 29–38.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology, 46*, 561–584. doi:10.1146/annurev.ps.46.020195.003021

Shaffer, J. P. (2002). Multiplicity, directional (Type II) errors, and the null hypothesis. *Psychological Methods, 7*(3), 356–369. doi:10.1037/1082-989X.7.3.356

Shaffer, J. P., Kowalchuk, R. K., & Keselman, H. J. (2013). Error, power, and cluster separation rates of pairwise multiple testing procedures. *Psychological Methods, 18*(3), 352–367. doi:10.1037/a0032478

Share, D. L. (1984). Interpreting the output of multivariate analyses: A discussion of current approaches. *British Journal of Psychology*, *75*(3), 349–362. doi:10.1111/j.2044-8295.1984.tb01905.x

Sheskin, D. J. (2007). *Handbook of parametric and nonparametric statistical procedures* (4th ed.). London, U.K.: Chapman & Hall.

Sheskin, D. J. (2011). *Handbook of parametric and nonparametric statistical procedures* (5th ed.). Boca Raton, FL: Chapman & Hall.

Shulz, K. F ., & Grimes, D. A. (2005). Multiplicity in randomised trials I: Endpoints and treatments. *The Lancet, 365*(9470), 1591–1595. doi:10.1016/S0140-6736(05)66461-6

Shulz, K. F., & Grimes, D. A. (2006). *The Lancet handbook of essential concepts in clinical research*. Edinburgh, U.K.: Elsevier.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distrubutions. *Journal of the American Statistical Association, 62*(318), 626–633. doi:10.1080/01621459.1967.10482935

Sijbrandij, M., Engelhard, I. M., Lommen, M. J. J., Leer, A., & Baas, J. M. P. (2013). Impaired fear inhibition learning predicts the persistence of symptoms of posttraumatic stress disorder (PTSD). *Journal of Psychiatric Research, 47*(12), 1991–1997. doi:10.1016/j.jpsychires.2013.09.008

Simes, R. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, *73*(3), 751–754. doi:10.1093/biomet/73.3.751

Sinclair, J. K., Taylor, P. J., & Hobbs, S. J. (2013). Alpha level adjustments for multiple dependent variable analyses and their applicability—a review. *International Journal of Sports Science and Engineering, 7*(1), 17–20.

Sirkin, R. M. (2006). *Statistics for the social sciences* (2nd ed.). Thousand Oaks, CA: Sage.

Small, D. S., Volpp, K. G., & Rosenbaum, P. R. (2011). Structured testing of 2×2 factorial effects: An analytic plan requiring fewer observations. *The American Statistician, 65*(1), 11–15. doi:10.1198/tast.2011.10130

Smyth, B. P., James, P., Cullen, W., & Darker, C. (2015). "So prohibition can work?" Changes in the use of novel psychoactive substances among adolescents attending a drug and alcohol treatment service following a legislative ban. *The International Journal of Drug Policy, 26*(9), 887–889. doi:10.1016/j.drugpo.2015.05.021

Spector, P. E. (1981). Multivariate data analysis for outcome studies. *American Journal for Community Psychology, 9*(1), 45–53. doi:10.1007/BF00896359

Spicer, J. (2005). *Making sense of multivariate data analysis*. Thousand Oaks, CA: Sage.

Springer. (2016). Open access review of "A randomized, controlled trial of dietary improvement for adults with major depression (the 'SMILES' trial)". Retrieved June 30, 2019, from https://static-content.springer.com/openpeerreview/art%3A10.1186%2Fs12916-017-0791-y/12916_2017_791_ReviewerReport_V1_R2.pdf

Stangor, C. (2015). *Research methods for the behavioral sciences* (5th ed.). Stamford, CT: Cengage.

Stanley, J. C. (1957). Additional "post-mortem" tests of experimental comparisons. *Psychological Bulletin, 54*(2), 128–130. doi:10.1037/h0041617

Steinfatt, T. M. (2006). The alpha percentage and experimentwise error rates in communication research. *Human Communication Research, 5*(4), 366–374. doi:10.1111/j.1468-2958.1979.tb00650.x

Stenfors, C. U. D., Marklund, P., Hanson, L. L. M., Theorell, T., & Nilsson, L. (2014). Are subjective cognitive complaints related to memory functioning in the working population? *BMC Psychology, 2*(3). doi:10.1186/2050-7283-2-3

Stephenson, A. G. (2002). evd: extreme value distributions. *R News, 2*(2), 31–32. Retrieved June 30, 2019, from https://cran.r-project.org/doc/Rnews/

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Lawrence Erlbaum.

Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). New York, NY: Taylor & Francis.

Strassburger, K., & Bretz, F. (2008). Compatible simultaneous lower confidence bounds for the Holm procedure and other Bonferroni based closed tests. *Statistics in Medicine, 27*(24), 4914–4927. doi:10.1002/sim.3338

Strahan, R.F. (1982). Multivariate analysis and the problem of Type I error. *Journal of Counseling Psychology, 29*(2), 175–179. doi:10.1037/0022-0167.29.2.175

Streiner, D. L. (2015). Best (but oft-forgotten) practices: The multiple problems of multiplicity—whether and how to correct for many statistical tests. *American Journal of Clinical Nutrition, 102*(4), 721–728. doi:10.3945/ajcn.115.113548

Streiner, D. L, & Norman, G. R. (2011). Correction for multiple testing: Is there a resolution? *Chest, 140*(1), 16–18. doi:10.1378/chest.11-0523

Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.

Tamhane, A. C. (2009). *Statistical analysis of designed experiments: Theory and applications*, Hoboken, NJ: John Wiley & Sons.

Tang, J., Fullarton, R., Samson, S.-L., & Chen, Y. (2019). Delayed cord clamping does not affect umbilical cord blood gas analysis. *Archives of Gynecology and Obstetrics, 299*(3), 719–724. doi:0.1007/s00404-019-05048-5

Thompson, B. (1994). Planned versus unplanned and orthogonal versus nonorthogonal contrasts: The neo-classical perspective. In B. Thompson (Ed.), *Advances in social science methodology* (vol. 3; pp. 3–27). Greenwich, CT: JAI Press.

Toothaker, L. E. (1992). *Multiple comparison procedures*. Thousand Oaks, CA: Sage.

Tromovitch, P. (2012). Statistical reporting with Philip's sextuple and extended sextuple: A simple method for easy communication of findings. *Journal of Research Practice, 8*(1), Article P2. Retrieved June 30, 2019, from http://jrp.icaap.org/index.php/jrp/article/view/323/270

Tucker, M. L. (1991). A compendium of textbook views on planned versus post hoc tests. In B. Thompson (Ed.), *Advances in education research: Substantive findings, methodological developments* (vol. 1; pp. 107–118). Greenwich, CT: JAI Press.

Tukey, J. W. (1953). The problem of multiple comparisons. In H. I. Braun (Ed.), *The collected works of John W. Tukey VIII. Multiple comparisons: 1948–1983* (pp. 1–300). New York, NY: Chapman & Hall.

Turner, G. R., Novakovic-Agopian, T., Kornblith, E., Adnan, A., Madore, M., Chen, A. J. W., & D'Esposito, M. (2019). Goal-oriented attention self-regulation (GOALS) training in older adults. *Aging & Mental Health*. doi:10.1080/13607863.2018.1534080

U.S. Department of Health and Human Services. (1998). *Guidance for industry: E9 statistical principles for clinical trials*. Retrieved June 30, 2019, from https://www.fda.gov/media/71336/download

van der Velden, P. G., Setti, I., van der Meulen, E., & Das, M. (2019). Does social networking sites use predict mental health and sleep problems when prior problems and loneliness are taken into account? A Population-based prospective study. *Computers in Human Behavior, 93*, 200–209. doi:10.1016/j.chb.2018.11.047

van Gils, E. J. M., Veenhoven, R. H., & Hak, E. (2009). Effect of reduced-dose schedules with 7-valent pneumococcal conjugate vaccine on nasopharyngeal pneumococcal carriage in children: A randomized controlled trial. *JAMA, 302*(2), 159–167. doi:10.1001/jama.2009.975

Van Patten, R., Greif, T., Britton, K., & Tremont, G. (2019). Single-photon emission computed tomography (SPECT) perfusion and neuropsychological performance in mild cognitive impairment. *Journal of Clinical and Experimental Neuropsychology, 41*(5), 530–543. doi:10.1080/13803395.2019.1586838

Wang, L. (1993, November). Planned versus unplanned contrasts: Exactly why planned contrasts tend to have more power against Type II error. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, New Orleans, LA. Retrieved June 30, 2019, from http://files.eric.ed.gov/fulltext/ED364598.pdf

Weber, D. C., & Skillings, J. H. (2000). *A first course in the design of experiments*. Boca Raton, FL: CRC.

Weiss, D. J. (2006). *Analysis of variance and functional measurement*. New York, NY: Oxford University Press.

Weisse, K., Winkler, S., Hirche, F., Herberth, G., Hinz, D., Bauer, M., Röder, S., Rolle-Kampczyk, U., von Bergen, M., Olek, S., Sack, U., Richter, T., Diez, U., Borte, M., Stangl, G. I., & Lehmann, I. (2013). Maternal and newborn vitamin D status and its impact on food allergy development in the German LINA cohort study. *Allergy, 68*(2), 220–228. doi:10.1111/all.12081

Welkowitz, J., Cohen, B. H., & Lea, R. B. (2012). *Introductory statistics for the behavioral sciences* (7th ed.). Hoboken, NJ: John Wiley & Sons.

Westfall, P. (1985). Simultaneous small-sample multivariate Bernoulli confidence intervals. *Biometrics, 41*(4), 1001–1013. doi:10.2307/2530971

Westfall, P. H., Tobias, R. D., & Wolfinger, R. D. (2011). *Multiple comparisons and multiple tests using SAS* (2nd ed.). Cary, NC: SAS Institute.

Wetcher–Hendricks, D. (2011). *Analyzing quantitative data*. Hoboken, NJ: John Wiley & Sons.

Williams, J. D. (1973). Note on familywise error rates. *Psychological Reports, 32*, 1221–1222. doi:10.2466/pr0.1973.32.3c.1221

Woodward, J. A., Bonett, D. G., & Brecht, M. (1990). *Introduction to linear models and experimental design*. San Diego, CA: Harcourt Brace Jovanovich.

Wright, S. P. (1992). Adjusted *p*-values for simultaneous inference. *Biometrics, 48*, 1005–1013. doi:10.2307/2532694

Young, S. S. (2008). *Everything is dangerous: A controversy*. Paper presented at the RAND Statistics Seminar, Pittsburgh, PA. Retrieved June 30, 2019, from https://www.niss.org/sites/default/files/Young_Safety_June_2008.pdf

Zieffler, A. S., Harring, J. R., & Long, J. D. (2011). *Comparing groups: Randomization and bootstrap methods using R*. Hoboken, NJ: John Wiley & Sons.

Zintzaras, E., & Lau, J. (2008). Synthesis of genetic association studies for pertinent gene–disease associations requires appropriate methodological and statistical approaches. *Journal of Clinical Epidemiology, 61*(7), 634–645. doi:10.1016/j.jclinepi.2007.12.011

Zwick, R. (1986). Testing pairwise contrasts in one-way analysis of variance designs. *Psychoneuroendocrinology, 11*(3), 253–276. doi:10.1016/0306-4530(86)90013-2