Journal of Research Practice
Volume 11, Issue 1, Article P2, 2015

ISSN: 1712-851X

Provocative Idea:

Planned Hypothesis Tests Are Not Necessarily Exempt From Multiplicity Adjustment

Andrew V. Frane
Department of Psychology
University of California, Los Angeles
UNITED STATES
avfrane@ucla.edu

Abstract

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. I critically evaluate this 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.

Index Terms: hypothesis testing; null hypothesis; statistical inference; statistical methods; hypothesis testing; Type I error

Suggested Citation: Frane, A. V. (2015). Planned hypothesis tests are not necessarily exempt from multiplicity adjustment. *Journal of Research Practice*, *11*(1), Article P2. Retrieved from http://jrp.icaap.org/index.php/jrp/article/view/514/417

1. Background

1.1. 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 observed in the sample actually applies to the population or whether it is 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 frequently the goal of research is 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), 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.

1.2. 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, for example, when evaluating more than one mean difference or more than one correlation. The resulting problem of *multiplicity* (multiple testing) is known to many researchers: 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 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* (e.g., the correlation) 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 estimated *effect sizes*, rather than merely about whether the effects are zero or nonzero. Note that although this article generally discusses multiplicity in terms of null hypothesis testing, the same principles of multiplicity are relevant to computing confidence intervals for effect size estimates.

1.3. 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, presented here 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 can inflate as the number of tests increases.

1.3.1. Per-Family Type I Error Rate (PFER)

The PFER (Tukey, 1953) 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).

1.3.2. Familywise Type I Error Rate (FWER)

The FWER (Tukey, 1953) 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 methods of FWER control (e.g., Holm, 1979; Hommel, 1988) can provide more statistical power; see Dmitrienko, Tamhane, and Bretz (2010) for a litany of such methods, 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), it has very limited applicability (Benjamini, 2010; Goeman & Solari, 2014; Hsu, 1996; Tamhane, 2009). In fact, most methods of Type I error control do not require omnibus tests at all (Dmitrienko, Tamhane, & Bretz, 2010).

1.3.3. False Discovery Rate (FDR)

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 (Benjamini & Hochberg, 1995). Roughly 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, Tamhane, & Bretz, 2010). Note also that the relevance of the FDR is limited when hypotheses have unequal likelihoods, because tests that are known to produce low *p*-values (call them "ringers") can drive down the FDR, thereby allowing tests with higher *p*-values to become statistically significant (Finner & Roters, 2001).

1.4. 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," for example, 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, 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" (p. 305). 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 scientific harm. Hence, multiplicity adjustments should account for each opportunity for scientific harm, that is, each test that would constitute a discovery on its own if statistically significant. The number of potential discoveries in a given study is often straightforward, but sometimes 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 gender: 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 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, that is, 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 gender 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 for gender could make an interesting refinement of the results, and a method-gender interaction could be a relevant caveat to the results, but only a main effect for teaching method has the potential to generate approval for the new method (in fact, a method-gender interaction might even *prevent* approval).

Clearly, the potential for 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.

2. Planned-Hypotheses Exemption From Multiplicity Adjustment (PHEMA)

As numerous authors have noted (e.g., Anderson, 2014; Glickman, Rao, & Shultz, 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 article 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 an empirical justification, it does come with a seductive offer: more statistical power.

Table 1. *Defense of Unadjusted Multiple Testing*

Study	Journal	Excerpt
Cachelin et al., 2014, p. 453	*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, p. 257	*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, p. 39	*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, p. 3	*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, p. 8	*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, p. 103	*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, Metz, & Pakulski, 2014, p. 253	*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, p. 1133	*Journal of General Internal Medicine*	"All analyses were planned a priori; therefore, p values were not adjusted for multiple comparisons."
Stenfors et al., 2014, p. 5	*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."

3. 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 from the original research questions. Planned comparisons are distinguished from *unplanned comparisons*, which are performed without any a priori expectation, for example, 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, Cohen, West, & Aiken, 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 conducted following 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).

4. Rationalizations for PHEMA

4.1. 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.

4.2. 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?

5. 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, for example, 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 (as per Google Scholar). One such article 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 levelled at Bonferroni and other corrections (e.g., Perneger, 1998), the analyses were performed without adjustment. (Roche & Chainay, 2013, p. 1017)

Sijbrandij, Engelhard, Lommen, Leer, & Baas (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)" (p. 1993). For other PHEMA-based statements citing Perneger, see Askari, Kirby, Parker, Thompson, & O'Neill (2013), Clifford et al. (2012), Fenesi, Heisz, Savage, Shore, & Kim (2014), Kawai et al. (2014), Krane-Gartiser, Henriksen, Morken, Vaaler, & Fasmer (2014), Lau, Lin, & Flores (2012), Weisse et al. (2013), and many others.

6. Variations on PHEMA

6.1. 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, Harring, & Long, 2011). In fact, some researchers have explicitly defended their unadjusted comparisons on that basis (e.g., Harkness & Luther, 2001; Nam & Zellner, 2011; Nieuwenhuis, Folia, Forkstam, Jensen, & Petersson, 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 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 is generally not plausible for two-sided tests.

6.2. 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, p. 7).

For multigroup 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). 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, Cohen, & Lea, 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?

6.3. Reverse-PHEMA

Some authors have proposed the opposite of PHEMA: that planned tests require multiplicity adjustment and that unplanned tests are exempt (e.g., Rovai, Baker, & Ponton, 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.

7. 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 (European Agency for the Evaluation of Medicinal Products, 2002; U.S. Department of Health and Human Services, 1998). 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 article 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.

7.1. Recommendations for Researchers

(a) 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.

(b) 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 correct discoveries (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.

(c) As recommended by the American Psychological Association (2012) and by other sources (including a previous article in this journal; 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.

(d) 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. 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."

7.2. Recommendations for Professors and Textbook Authors

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

(b) 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.

(c) 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 scientific harm is caused by false discoveries, and the purpose of multiplicity adjustments is to preserve that limit when there are multiple simultaneous opportunities for scientific 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.

References

Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Aickin, M. (1999). Other method for adjustment of multiple testing exists. *BMJ, 318*(7176), 127-128.

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, D.C.: Author.

Anderson, N. H. (2014). *Empirical direction in design and analysis*. New York, NY: Routledge.

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optics, 34*(5), 502-508.

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.

Baguley, T. S. (2012). *Serious stats: A guide to advanced statistics*. New York, NY: Palgrave Macmillan.

Bechofer, R. E., & Dunnett, C. W. (1982). Multiple comparisons for orthogonal contrasts: Examples and tables. *Technometrics, 24*(3), 213-222.

Bender, R., & Lange, S. (1998). What's wrong with arguments against multiplicity adjustments [Letter to the editor]. *BMJ*. Retrieved 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.

Benjamini, Y. (2010). Simultaneous and selective inference: Current successes and future challenges. *Biometrical Journal, 52*(6), 708-721.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B,57*(1), 289-300.

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.

Bretz, F., & Westfall, P. H. (2014). Multiplicity and replicability: Two sides of the same coin.*Pharmaceutical Statistics, 13*(6), 343-344.

Brown, S. R. (1990). *Experimental design and analysis*. Newbury Park, CA: Sage.

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.

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.

Cohen, B. H. (2013). *Explaining psychological statistics* (4th ed.). Hoboken, NJ: John Wiley.

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.

Dmitrienko, A., Tamhane, A. C., & Bretz, F. (Eds.). (2010). *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, UK: Cambridge University Press.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association, 56*(293), 52-64.

European Agency for the Evaluation of Medicinal Products. (2002). *Points to consider on multiplicity issues in clinical trials*. Retrieved from http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003640.pdf

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.

Finner, H., & Roters, M. (2001). On the false discovery rate and expected Type I errors. *Biometrical Journal*, *43*(8), 985-1005.

Glaus, J., Vandeleur, C. L., von Känel, R., Lasserre, A. M., Strippoli, M. F., Gholam-Rezaee, 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.

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.

Goeman, J. J., & Solari, A. (2014). Multiple hypothesis testing in genomics. *Statistics in Medicine*, *33*(11), 1946-1978.

Ha, R. R., & Ha, J. C. (2012). *Integrative statistics for the social and behavioral sciences*. Thousand Oaks, CA: Sage.

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.

Hays, W. L. (1988). *Statistics* (4th ed.). Fort Worth, TX: Holt, Rinehart and Winston.

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.

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.

Hommel, G. (1988). A stagewise multiple test procedure based on a modified Bonferroni test. *Biometrika*, *75*(2), 383-386.

Howell, D. C. (2013). *Statistical methods for psychology* (8th ed.). Belmont, CA: Wadsworth.

Hsu, J. C. (1996). *Multiple comparisons*. Boca Raton, FL: Chapman & Hall.

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), 5-9.

Kawai, V. K., Avalos, I., Oeser, A., Oates, J. A., Milne, G. L., Solus, J., . . . 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.

Keppel, G., & Zedeck, S. (1989). *Data analysis for research designs*. New York, NY: W. H. Freeman.

Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences* (4th ed.). Thousand Oaks, CA: Sage.

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.

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.

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.

McHugh, R. B., & Ellis, D. S. (1957). The 'post-mortem' testing of experimental comparisons. *Psychological Bulletin*, *52*(5), 425-428.

McKillup, S. (2012). *Statistics explained: An introductory guide for life scientists*. Cambridge, UK: Cambridge University Press.

Miller, R. G. (1981). *Simultaneous statistical inference* (2nd ed.). New York, NY: Springer-Verlag.

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.

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.

O'Keefe, D. J. (2003). Colloquy: Should familywise alpha be adjusted? Against familywise alpha level adjustment. *Human Communication Research*, *29*(3), 431-447.

Pagano, R. R. (2013). *Understanding statistics in the behavioral sciences* (10th ed.). Boston, MA: Cengage.

Pashler, H., & Wagenmakers, E. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528-530. Retrieved from http://pps.sagepub.com/content/7/6/528.full.pdf

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.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design and analysis: An integrated approach*. New York, NY: Psychology.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *BMJ*, *316*(7139), 1236-1238.

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.

Randolph, K. A., & Meyers, L. L. (2013). *Basic statistics in multivariate analysis*. Oxford, England: Oxford University Press.

Roche, K., & Chainay, H. (2013). Visually guided grasping of common objects: Effects of priming. *Visual Cognition*, *21*(8), 1010-1032.

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.

Rutherford, A. (2011). *ANOVA and ANCOVA: A GLM approach*. Hoboken, NJ: John Wiley.

Ryan, T. A. (1959). Multiple comparisons in psychological research. *Psychological Bulletin*,*56*(1), 26-47.

Ryan, T. A. (1962). The experiment as the unit for computing rates of error. *Psychological Bulletin*, *59*(4), 301-305.

Ryan, T. A. (1995, December 1). 'Post hoc' tests [posted to STAT-L discussion group]. Retrieved from http://groups.google.com/forum/#!activity/sci.stat.consult/JznJOuHfr7QJ/sci.stat.consult/fJmYD9TJQ6A/AfMk-8gvJacJ

Sheskin, D. (2011). *Handbook of parametric and nonparametric statistical procedures* (5th ed.). Boca Raton, FL: Chapman & Hall.

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.

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.

Steinfatt, T. M. (2006). The alpha percentage and experimentwise error rates in communication research. *Human Communication Research, 5*(4), 366-374.

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), 1-14. Retrieved from http://www.biomedcentral.com/content/pdf/2050-7283-2-3.pdf

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*. Advance online publication.

Streiner, D. L, & Norman, G. R. (2011). Correction for multiple testing: Is there a resolution? *Chest*, *140*(1), 16-18.

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.

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.

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 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. Braun (Ed.), *The collected works of John W. Tukey, Vol VIII: Multiple comparisons: 1948-1983* (pp. 1-300). New York, NY: Chapman & Hall.

U.S. Department of Health and Human Services. (1998). *Guidance for industry: E9 statistical principles for clinical trials*. Retrieved from http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm073137.pdf

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 from http://files.eric.ed.gov/fulltext/ED364598.pdf

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, . . . 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.

Welkowitz, J., Cohen, B. H., & Lea, R. B. (2012). *Introductory statistics for the behavioral sciences* (7th ed.). Hoboken, NJ: John Wiley.

Zieffler, A. S., Harring, J. R., & Long, J. D. (2011). *Comparing groups: Randomization and bootstrap methods using R*. Hoboken, NJ: John Wiley.

Received 4 September 2015 | Accepted 9 October 2015 | Published 13 October 2015