5-1-2015

Per Family or Familywise Type I Error Control: "Eether, Eyether, Neether, Nyther, Let's Call the Whole Thing Off!"

H. J. Keselman
University of Manitoba, kesel@ms.umanitoba.ca

Invited Debate
Per Family or Familywise Type I Error Control: "Eether, Eyether, Neether, Nyther, Let's Call the Whole Thing Off!"[1]

H. J. Keselman
University of Manitoba
Winnipeg, Manitoba

Frane (2015) pointed out the difference between per-family and familywise Type I error control and how different multiple comparison procedures control one method but not necessarily the other. He then went on to demonstrate in the context of a two group multivariate design containing different numbers of dependent variables and correlations between variables how the per-family rate inflates beyond the level of significance. In this article I reintroduce other newer better methods of Type I error control. These newer methods provide more power to detect effects than the per-family and familywise techniques of control yet maintain the overall rate of Type I error at a chosen level of significance. In particular, I discuss the False Discovery Rate due to Benjamini and Hochberg (1995) and k-Familywise Type I error control enumerated by Lehmann and Romano (2005), Romano and Shaikh (2006), and Sarkar (2008). I conclude the article by referring readers to articles by Keselman, et al. (2011, 2012) which presented R computer code for determining critical significance levels for these newer methods of Type I error control.

Keywords: Type I error, multiple comparisons, simultaneous inference

Introduction

Frane (2015) presented an article which clarified the difference between the per-family (PFER) and familywise (FWER) Type I error rates (See also Klockars & Hancock, 1994). It is important that applied researchers understand the difference between the rates and how different multiple comparison procedures may control

---

[1] From the film "Shall We Dance?" Words by Ira Gershwin; music by George Gershwin. Introduced by Fred Astaire and Ginger Rogers.

H. J. Keselman is a Professor of Psychology and Associate Editor Emeritus of this journal. Email him at kesel@ad.umanitoba.ca.

one rate of error but not the other. For example, as he notes, the typical Dunn (1961)-Bonferroni method controls the overall rate of Type I error per-family, whereas other Bonferroni methods of Type I error control (e.g., Holm, 1979) control the familywise rate of error. Through simulation methods he then shows that in a multivariate design containing two groups, multiple dependent measures, and various correlations between the dependent variables, the FWER may be controlled, yet the PFER can be very large. The author also notes in the article that other issues could have been discussed such as newer methods of controlling Type I errors and other multiple comparison procedures themselves; some issues were noted but not discussed in detail.

My intention in this article is to take the reader further into the topics of Type I error control and multiple comparison procedures that Frane (2015) did not have the space to discuss. I believe these additional topics are very important to discuss since the issue of Type I error control has advanced immeasurably since the early discussions related to PFER and FWER control.

Per-experiment and experimentwise Type I error control

At the outset I want to expand on the definitions of per-family and familywise presented by Frane (2015). But first, I want to re-introduce the per-experiment (PEER) and the experimentwise (EWER) Type I error rates, rates applied researchers are more likely to be familiar with. Ryan (1959, 1960, 1962) in his seminal articles regarding overall Type I error control versus comparisonwise (CWE) (i.e., per test or per comparison) control, used the terminology per-experiment and experimentwise to indicate that these rates applied to controlling the maximum overall rate of Type I error for multiple tests of significance assessed within an experiment. Later in the history of methods for controlling the overall rate of Type I error, per-family and familywise became equated with per-experiment and experimentwise (See Hochberg & Tamhane, 1987).

The distinction is important because it allows one to adopt per-family and familywise control in more interesting and dynamic ways. For example, in a one-way design where a researcher computes pairwise and complex comparisons between group means, one can set a per-family or familywise error rate over each family of tests (i.e., the pairwise tests and complex comparisons tests), and thus maintain the per-experiment or experimentwise rates at some overall maximum value. So a .05 level of significance can be tied to each family of tests and consequently the maximum overall joint per-experiment or experimentwise probability of Type I error can be fixed at .10. To further illustrate the nuances of

familywise and experimentwise control consider an $A \times B$ design. In such a design a researcher can set familywise rates of error over all tests performed on the A effect, B effect, and $A \times B$ effects. Collectively, the overall or experimentwise Type I error rate would be a function of the three familywise rates. For example, suppose the researcher chose to perform all possible pairwise comparisons on the A main effect, a number of complex comparisons on the B main effect, and a number of interaction contrasts on the $A \times B$ effects setting a .05 value on each set. Collectively therefore, the overall experimentwise Type I error rate would be controlled at the .15 level. Clearly by thinking about the familywise or per-family rate as rates for related families of tests, the researcher can see the flexibility that s/he is afforded. I will have more to say on how researchers should define a family shortly.

Newer definitions of Type I error control

Background

Multiplicity of testing. The multiplicity problem in statistical inference refers to selecting the statistically significant findings from a large set of findings (tests) to either support or refute one's research hypotheses. Discussions on how to deal with multiplicity of testing have permeated many literatures for decades. There are those who believe that the occurrence of any false positive must be guarded at all costs (see Games, 1971; Ryan, 1960, 1962; Westfall & Young, 1993). That is, as promulgated by Thomas Ryan, pursuing a false lead can result in the waste of much time and expense, and is an error of inference that accordingly should be stringently controlled. Those in this camp deal with the multiplicity issue by setting α for the entire set of tests computed. This type of control has been referred to in the literature as experimentwise (EWER) or familywise (FWER) control. Those in the opposing camp maintain that stringent Type I error control results in a loss of statistical power and consequently important treatment effects go undetected (see Rothman, 1990; Saville, 1990). Members of this camp typically believe the error rate should be set per comparison [the probability of rejecting a given comparison] (the CWE rate) and usually recommend a five percent level of significance, allowing the overall error rate (i.e., EWER or FWER) to inflate with the number of tests computed. In effect, those who adopt comparisonwise control ignore the multiplicity issue.

Family size. Specifying family size is a very important component of multiple testing. As Westfall et al. (1999, p. 10) note, differences in conclusions reached from statistical analyses that control for multiplicity of testing (FWER) and those that do not (CWE) are directly related to family size. Specifically, the larger the family size, the less likely individual tests will be found to be statistically significant with FWER control. Accordingly, to achieve as much sensitivity as possible to detect true differences and yet maintain control over multiplicity effects, Westfall et al. recommend that researchers "choose smaller, more focused families rather than broad ones, and (to avoid cheating) that such determination must be made *a priori...*" (p. 10).

Not only does the FWER rate depend on the number of null hypotheses that are true but as well on the distributional characteristics of the data and the correlations among the test statistics. Because of this, an assortment of multiple comparison procedures have been developed, each intended to provide FWER control.

As I indicated at the outset, since the per-family/per-experiment and familywise/experimentwise error rates were introduced, researchers have defined new ways of controlling Type I errors which by-in-large are intended to provide control over multiple tests of significance that one does not achieve with comparisonwise control and more power to detect effects than is provided by the familywise and experimentwise rates.

The false discovery rate (FDR)

It was noted by Frane (2015) that this is a new definition of Type I error control that affords the user more power to detect true effects though at the cost of allowing a greater number of Type I errors. However, Frane believes that if researchers want more power they should exert better experimental control and/or use more subjects in their studies. Presuming that applied researchers are always attuned to controlling extraneous variance and accordingly adopt the best experimental control that is feasible for their studies, the remaining avenue to increase power to detect effects is to increase the number of participants examined in their studies. Not always however, possible. In my department the subject pool is limited and experimenters do not have access to as many subjects that comprise the pool. Thus, achieving more statistical power through more liberal definitions of Type I error control and more sensitive multiple comparison procedures should be a viable option for researchers to consider.

As indicated, several different error rates have been proposed in the multiple comparison literature. The majority of discussion in the literature has focused on the FWER, although other error rates, such as the FDR also have been proposed (e.g., Benjamini & Hochberg, 1995). The FDR is defined by these authors as the expected proportion of the number of erroneous rejections to the total number of rejections.

Use of the false discovery rate criterion has become widespread when making inferences in research involving the human genome, where family sizes in the thousands are common. See the review by Dudoit, Shaffer and Boldrick (2003), and references contained therein. Another area of research where FDR controlling procedures have had a significant impact is functional magnetic resonance imaging. In these experiments researchers are conducting numerous (often more than 100,000) significance tests that relate to tests of activation on specific voxels (i.e., areas) within the brain (e.g., Callan, Jones, Munhall, Callan, Kroos, & Vatikiotis-Bateson, 2003).

The Benjamini and Hochberg (1995) procedure has been shown to control the FWER for several situations of dependent tests, that is, for a wide variety of multivariate distributions that make their procedure applicable to most testing situations scientists might encounter (see Sarkar, 1998; Sarkar & Chang, 1997). In addition, simulation studies comparing the power of the Benjamini and Hochberg procedure to several FWER controlling procedures have shown that as the number of treatment groups increases (beyond 4 treatment groups), the power advantage of their procedure over the FWER controlling procedures becomes increasingly large (Keselman et al., 1999). The power of FWER controlling procedures is highly dependent on the family size (i.e., number of comparisons), decreasing rapidly with larger families (Holland & Cheung, 2002; Miller, 1981). Therefore, control of the FDR results in more power than FWER controlling procedures in experiments with many treatment groups, but yet provides more control over Type I errors than CWE controlling procedures.

Suppose for n means, $\mu_1, \mu_2, \ldots, \mu_J$, and our interest is in testing the family of $m = [J(J-1)]/2$ pairwise hypotheses, $H_0 : \mu_i - \mu_j = 0$, of which m_0 are true. Let S equal the number of correctly rejected hypotheses from the set of R rejections; the number of falsely rejected pairs will be V. In terms of the random variable V, the CWE is $E(V/m)$, while the FWER is given by $P(V \geq 1)$. Thus, testing each and every comparison at α guarantees that $E(V/m) \leq \alpha$, while according to the Bonferroni inequality, testing each and every comparison at level α/m guarantees that $P(V \geq 1) \leq \alpha$.

According to Benjamini and Hochberg (1995) the proportion of errors committed by falsely rejecting null hypotheses can be expressed through the random variable $Q = V / R$, that is, the proportion of rejected hypotheses that are erroneously rejected. (It is important to note that Q is defined to be zero when $R = 0$; that is, the error rate is zero when there are no rejections.) The FDR was defined by Benjamini and Hochberg as the mean of Q, that is

$$E(Q) = E\left(\frac{V}{R}\right), \text{ or } E(Q) = E\left(\frac{\text{Number of false rejections}}{\text{Number of rejections}}\right).$$

That is, the FDR is the expected proportion of false discoveries or false positives.

As Benjamini and Hochberg (1995) indicate, this error rate has a number of important properties:

a)   If $\mu_1 = \mu_2 = \cdots = \mu_J$, then all m (pairwise) comparisons truly equal zero, and therefore the FDR is equivalent to the FWER; that is, in the case of the complete null being true, FDR control implies FWER control. Specifically, in the case of the complete null hypothesis being true, $S = 0$ and therefore $V = R$. So, if $V = 0$, then $Q = 0$, and if $V > 0$ then $Q = 1$ and accordingly $P(V \geq 1) = E(Q)$.

b)   In testing the family of (pairwise) hypotheses, of which m_0 are true, when $m_0 < m$, the FDR is smaller than or equal to the FWER. The FDR is smaller than or equal to the FWER because in this case FWER $= P(R \geq 1) \geq E(V / R) = E(Q)$. This indicates that if the FWER is controlled for a procedure, then the FDR is as well. Moreover, if one adopts a procedure that provides FDR control, rather than strong (i.e., over all possible mean configurations) FWER control, then based on the preceding relationship, a gain in power can be expected.

c) V / R tends to be smaller when there are fewer pairs of equal means and when the non-equal pairs are more divergent, resulting in a greater differences in the FDR and the FWER values and thus a greater likelihood of increased power by adopting FDR control.

With the BH FDR procedure, the p-values corresponding to the m (pairwise) statistics for testing the hypotheses $H_1, H_2, \ldots, H_m$ are ordered from smallest to

largest, that is, $p_1 \leq p_2 \leq \cdots \leq p_m$. Let k be the largest value of i for which $p_i \leq (i/m)\alpha$ and then reject all H_i, $i = 1, 2, \ldots, k$. On the basis of this procedure, one begins by assessing the largest p-value, p_m, and then proceeds to smaller p-values as long as $p_i > (i/m)\alpha$. Testing stops when $p_i \leq (k/m)\alpha$.

The k-FWER criterion and procedures for its control[2]

The classical approach for controlling Type I errors for a family of many (say m) hypothesis tests is FWER control. Once the family is defined, control of the FWER requires that

$$FWER \leq \alpha$$

for all configurations of true and false hypotheses. It is well known that for non-independent tests the probability (Pr) of making one or more Type I errors is

$$FWER = \text{Pr(One or more Type I errors for } m \text{ tests)} < 1 - (1 - \alpha)^m$$

Examples of procedures that control the overall rate of Type I error when many tests of hypotheses are examined are the single-stage Bonferroni procedures (e.g., Dunn, 1961) and stepwise Bonferroni procedures (Hochberg, 1988; Holm, 1979). However, when there are many hypotheses to be examined they can be deficient in power to detect non-null hypotheses. Indeed, when the size of the family of hypotheses to be tested becomes large, FWER becomes very restrictive and not very powerful at detecting false null hypotheses. For example, for m tests of significance, the single-stage Bonferroni level of significance would be α/m and when m is large detecting non-null effects will be difficult. As Lehmann & Romano (2005) note "control of the FWER at conventional levels becomes so stringent that individual departures from the hypothesis have little chance of being detected" (p. 1139).

Accordingly, Type I error control is not the only issue researchers must consider when testing a hypothesis or set of hypotheses. As in the case of testing a single hypothesis, researchers must also consider the ability of a procedure to detect departures from the hypothesis when they do occur (Lehmann & Romano, 2005, p. 1139). To address this issue, Lehmann & Romano, as well as others (See the references cited in Lehmann & Romano) developed the k-FWER method of

---

[2] Keselman et al. (2012) previously introduced these procedures to the psychological audience. Their article also includes the mathematical underpinnings of the procedures.

Type I error control. As they note, with a larger family of hypotheses, one might be willing to allow the possibility of falsely rejecting k true null hypotheses. With the possibility of falsely rejecting more than one, two, three, etc. null hypothesis(es), one obtains more power to detect false null hypotheses. Lehmann and Romano (2005) define k-FWER as the probability of rejecting at least k true null hypotheses.

$$k\text{-FWER} = \Pr\{\text{reject at least } k \text{ hypotheses H}_i \text{ with } i \in I(P)\}$$

Here $I(P)$ denotes the set of true null hypotheses when P is the true probability distribution. Control of the k-FWER requires that k-FWER $\leq \alpha$ for all P. When $k = 1$, then k-FWER reduces to 1-FWER or FWER which controls the probability of rejecting at least one true null hypothesis.

To help the reader to fully appreciate k-FWER, I note the following. Consider what it means to control 2-FWER instead of 1-FWER (or simply FWER) at $\alpha = .05$? This would be equivalent to specifying that the probability of 2 or more false rejections is controlled at .05, whereas FWER controls the probability of any (i.e., 1 or more) false rejections at .05. In essence, then, 2-FWER implicitly tolerates 1 false rejection and makes no explicit attempt to control the probability of its occurrence, unlike FWER which tolerates no false rejections at all. More generally, then, k-FWER tolerates $k - 1$ false rejections, but controls the probability of k or more false rejections at an $\alpha = .05$.

Before presenting these newer methods I provide some additional clarification of the k-FWER. First, remember that FWER control treats rejections of multiple true null hypotheses as being no more serious than the rejection of only one (i.e., at least one) true null hypothesis. The newer procedures have the same conceptual underpinning; however, for them falsely rejecting multiple true null hypotheses is no more serious than the rejection of only two, three, etc. true null hypotheses (i.e., at least 2, 3, etc.). Accordingly, a clean outcome from an analysis controlling the FWER is an outcome with no Type I errors. A clean outcome from a k-FWER analysis is an outcome with no more than $k - 1$ Type I errors. Note that in both cases, the number of Type I errors produced when at least k are produced (1 in the case of FWER) is of no concern as far as the error rate criterion is concerned.

Keselman, Miller and Holland (2011) describe four procedures that utilize the k-FWER method of multiple testing control. Technical descriptions can be

found in Keselman et al. (2011). As well these authors provide R code for running the newer procedures (See also Keselman et al., 2012).[3]

The Holm and generalized Holm (Lehmann and Romano) procedures

Lehmann and Romano (2005) provided a generalization of the Holm (1979) procedure. Just as the Holm procedure controls FWER under all dependency conditions, the generalized procedure controls k-FWER under the same dependency conditions (i.e., there are no dependency conditions).

The ordered p-values for the m individual tests denoted $p_{(1)} \leq \cdots \leq p_{(k)} \leq \cdots \leq p_{(m)}$ correspond to hypotheses, $H_{(1)}, \ldots, H_{(k)}, \ldots, H_{(m)}$. The generalized Holm procedure is defined stepwise as follows:

Step 0.   Let $i = 1$, k and α are chosen by the experimenter.

Step 1.   If $i \leq k$, go to step 2. If $k < i \leq m$, go to step 3. Otherwise, stop and reject all of the hypotheses.

Step 2.   If $p_{(i)} > \dfrac{k\alpha}{m}$, go to step 4. Otherwise, set $i = i + 1$ and go to step 1.

Step 3.   If $p_{(i)} > \dfrac{k\alpha}{m+k-i}$, go to step 4. Otherwise, set $i = i + 1$ and go to step 1.

Step 4.   Reject $H_{(j)}$ for $j < i$ and accept $H_{(j)}$ for $j \geq i$.

The Hochberg and generalized Hochberg (Sarkar 1) procedures

The generalization of the Hochberg (1988) procedure is a step up version of the generalized Holm procedure presented by Lehmann and Romano. Sarkar (2008) states that it controls k-FWER when the test statistics are independent or when they satisfy the multivariate totally positive order of two (MTP$_2$) condition.[4]

A step up procedure based on the same set of critical values as a step down procedure will always reject at least as many hypotheses and therefore will be

---

[3] The R code provides users with adjusted p-values. In its typical application, researchers compare a test statistic to a FWER critical value. Another approach for assessing statistical significance is with adjusted p-values, $\tilde{p}_i$, $i = 1, \ldots, m$ (Westfall et al., 1999; Westfall & Young, 1993). As Westfall and Young note "$\tilde{p}_i$ is the smallest significance level for which one still rejects a given hypothesis (H_i) in a family, given a particular (familywise) controlling procedure." (p. 11) The advantage of adjusted p-values for multiple comparison procedures, as with p-values for tests in comparisonwise contexts, is that they are more informative than merely declaring retain or reject H_i; they are a measure of the weight of evidence for or against the null hypothesis when controlling FWER. For example, if $\tilde{p}_i = 0.09$, the researcher/reader can conclude that the test is statistically significant at the FWER = 0.10 level, but not at the FWER = 0.05 level. Adjusted p-values are provided by the SAS system for many popular multiple comparison procedures (See Westfall et al., 1999). SPSS also provides adjusted p-values for most multiple comparison procedures.

[4] Keselman et al. (2012) define MTP$_2$ in their article.

more powerful at detecting false null hypotheses. I therefore recommend using the generalized Hochberg procedure over the generalized Holm procedure as long as the Hochberg procedure is appropriate to use.

The generalized Hochberg procedure is defined stepwise as follows:

Step 0.   Let $i = m$, k and α are chosen by the experimenter.

Step 1.   If $i > k$, go to step 2. If $1 \leq i \leq k$, go to step 3. Otherwise, stop and accept all of the hypotheses.

Step 2.   If $p_{(i)} \leq \dfrac{k\alpha}{m + k - i}$, go to step 4. Otherwise, set $i = i - 1$ and go to step 1.

Step 3.   If $p_{(i)} \leq \dfrac{k\alpha}{m}$, go to step 4. Otherwise, set $i = i - 1$ and go to step 1.

Step 4.   Reject $H_{(j)}$ for $j \leq i$ and accept $H_{(j)}$ for $j > i$.

Romano and Shaikh procedure   Romano and Shaikh (2006) developed a generalized version of the Hochberg procedure that has no dependency restrictions associated with it. This fact makes it attractive in situations with complex dependency conditions, i.e., such as when the family of tests are that the elements of a correlation matrix are zero. Step up tests such as the Hochberg are more powerful at detecting false null hypotheses than the step down test using the same critical values. However, since this generalized Hochberg test is valid to use under all dependency conditions, it does not use the same critical values as the generalized Holm procedure. The critical values are approximately halved. This negatively affects power to detect false null hypotheses since the *p*-values must be less than the critical values to be declared statistically significant. See Keselman et al.'s (2011) Appendix A for more information.

Sarkar 2 procedure   The Sarkar (2008) procedure is another generalized version of the Hochberg procedure. It controls *k*-FWER when the joint distribution of the *p*-values is multivariate totally positive of order two (MTP$_2$) in addition to having identical *k*th-order joint distributions under the null hypotheses. MTP$_2$ is a somewhat restrictive condition that is violated if any of the test statistics are negatively correlated, but met if the tests are pairwise independent (Sarkar, 2000). An example of a MTP$_2$ procedure would be many to one contrasts in a balanced design as is found in a Dunnett's one-sided comparisons with a control.

When the p-values are independent, this procedure has been found to be a more powerful generalized Hochberg procedure than a step up version of the generalized Holm procedure when $2 \leq k \leq 1 / \alpha$ (Sarkar, 2008). When $k = 1$, the Sarkar procedure is equivalent to the Hochberg procedure. Although, the Sarkar procedure is valid to use as long as the p-values have a MTP_2 distribution, we only recommend its use when the p-values are independent [See Keselman et al.'s (2011) Table 1 for a description of k-FWER method and type of dependency assumed to exist between the test statistics and associated p-values]. (Note: The R code provided in their Appendix B is only valid for the Sarkar procedure when the p-values are independent.)

Discussion

As the reader can see, the way in which Type I errors can be controlled for families of tests goes way beyond the per-family and familywise rates discussed by Frane (2015). The intention of my article was to review methods previously presented in the statistical and psychological literatures, with the intention of letting the reader see that researchers have many techniques that can be adopted to control the overall rate of Type I error. I recommend that applied researchers give serious consideration to the newer techniques (FDR and k-FWER) because they provide more power to detect non-null effects and yet limit the overall rate of Type I error at some specified value. So referring back to the title of this article I would say with regard to per-family or familywise control—eether, eyether, or perhaps neether, nyther.[5] The reader should note that the R code provided in Keselman et al. (2011, 2012) provides adjusted p-values for all of the newer methods discussed in this article. Users must select a method of control before cherry-picking the method that has the greatest number of statistically significant findings as reported through the R code.

---

[5] The methods described in this paper do not provide confidence intervals as compared to simultaneous MCPs [procedures that use one critical value to assess statistical significance such as Tukey's (1953) method]; they, nonetheless, should be considered an important tool in any data analyst's arsenal of viable methods for investigating treatment effects through many tests of significance.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, 57*(1), 289-300. doi:10.2307/2346101

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport*, *14*(17), 2213-2218.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*(293), 52-64. doi:10.1080/01621459.1961.10482090

Dudoit, S., Shaffer, J. P. & Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, *18*(1), 71-103. doi:10.1214/ss/1056397487

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).

Games, P. A. (1971). Multiple comparisons of means. *American Educational Research Journal*, *8,* 531-565. doi:10.3102/00028312008003531

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.

Holland, B. & Cheung, S. H. (2002). Familywise robustness criteria for multiple comparison procedures. *Journal of the Royal Statistical Society*, *B, 64*(1), 63-77. doi:10.1111/1467-9868.00325

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

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., Miller, C. E., & Holland, B. (2011). Many tests of significance: New methods for controlling Type I errors. *Psychological Methods*, *16*(4), 420-431. doi:10.1037/a0025810

Keselman, H. J., Miller, C. E., & Holland, B. (2012). Correction to many tests of significance: New methods for controlling Type I errors. *Psychological Methods*, 17(4), 679. doi:10.1037/a0030995

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

Lehmann, E. L., & Romano, J. P. (2005). Generalizations of the familywise error rate. *Annals of Statistics*, *33*(3), 1138–1154. doi:10.1214/009053605000000084

Miller, R. G. (1981). *Simultaneous statistical inference*. (2nd ed.) New York: McGraw-Hill.

Romano, J. P., & Shaikh, A. M. (2006). Stepup procedures for control of generalizations of the familywise error rate. *Annals of Statistics*, *34,* 1850–1873. doi:10.1214/009053606000000461

Rothman, K. (1990). No adjustments are needed for multiple comparisons. *Epidemiology*, *1*(1), 43-46.

Ryan, T. A. (1959). Multiple comparisons in psychological research. *Psychological Bulletin*, *56*(1), 26-47. doi:10.1037/h0042478

Ryan, T. A. (1960). Significance tests for multiple comparison proportions, variances and other statistics. *Psychological Bulletin*, *57*(4), 318-328. doi:10.1037/h0044320

Ryan, T. A. (1962). The experiment as the unit for computing rates of error. *Psychological Bulletin*, *59*(4), 301-305. doi:10.1037/h0040562

Sarkar, S. K. (1998). Some probability inequalities for ordered MTP2 random variables: A proof of the Simes conjecture. *Annals of Statistics*, *26*(2), 494–504. doi:10.1214/aos/1028144846

Sarkar, S. K., (2000). A note on the monotonicity of the critical values of a step–up test. *Journal of Statistical Planning Information, 87*(2), 241-249. doi:10.1016/S0378-3758(99)00200-1

Sarkar, S. K. (2008). Generalizing Simes' test and Hochberg's stepup procedure. *Annals of Statistics*, *36*(1), 337–363. doi:10.1214/009053607000000550

Sarkar, S. K., & Chang, C. K. (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association*, *92*(440), 1601–1608.

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

Tukey, J. W. (1953). The problem of multiple comparisons. In H. Braun (Ed.), *The collected works of John W. Tukey volume VIII, multiple comparisons: 1948-1983* (pp. 1-300). New York, NY: Chapman & Hall.

Westfall, P. H., Tobias, R. D., Rom, D., Wolfinger, R. D., & Hochberg, Y. (1999). *Multiple comparisons and multiple tests*. Cary, NC: SAS Institute, Inc.

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