Journal of Modern Applied Statistical Methods

Volume 14 | Issue 1 Article 5

5-1-2015

Are Per-Family Type I Error Rates Relevant in Social and Behavioral Science?

Andrew V. Frane
California State University - Los Angeles, avfrane@gmail.com

Are Per-Family Type I Error Rates Relevant in Social and Behavioral Science?

Cover Page Footnote
The author is grateful for the assistance of Theodore S. Bell.

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, 12-23.

ISSN 1538 − 9472

Invited Debate
Are Per-Family Type I Error Rates Relevant in Social and Behavioral Science?

Andrew V. Frane
University of California Los Angeles
Los Angeles, CA

The familywise Type I error rate is a familiar concept in hypothesis testing, whereas the per‑family Type I error rate is rarely addressed. This article uses Monte Carlo simulations and graphics to make a case for the relevance of the per‑family Type I error rate in research practice and pedagogy.

Keywords: Type I error, multiple comparisons, simultaneous inference

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 may stem largely from the fact that it is a stricter standard than the FWER, so controlling it can be more costly in statistical power (potentially increasing the Type II error rate). Given the tremendous pressure on researchers to find statistically significant *p*-values, any reduction in statistical power is a hard sell. However, as noted by a previous article in this journal (Barnette & McLean, 2005) and by others (Klockars & Hancock, 1994; Ryan, 1959, 1962), it is arguable that the PFER is often more relevant than the FWER in social and behavioral science research. 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

Mr. Frane is a doctoral student of cognitive psychology in the Visual and Multisensory Perception Lab. Email him at avfrane@ucla.edu.

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 article 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, Gamst, & Guarino, 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., 2010; Hsu, 1996).

This 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 correlated.

Controlling the PFER using the Bonferroni procedure

The Bonferroni procedure caps the maximum PFER at α by testing each hypothesis at a nominal 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 nominal 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. Like the Bonferroni procedure, Holm's method also allows 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, then tested sequentially in that order, each at a nominal 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., 2010) and is more powerful than Holm's, but generally does not allow computation of confidence intervals (Dmitrienko et al., 2010; 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 et al., 2010; 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., 2010) can be used when comparing multiple treatment groups to a placebo group. 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 confusion such as that 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, Glazko, & Yakovlev, 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 is not 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.

Psychological researchers that have touted the superior power of stepwise methods over the Bonferroni procedure (e.g., Blakesley et al., 2009; Eichstaedt, Kovatch, and Maroof, 2013; Seaman, Levin, & Serlin, 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). In fact, the sometimes dramatically 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". Ryan (1959, p. 40) called the choice between FWER and PFER "merely a matter of computational convenience." 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 demonstrations of analogous PFER inflation in other contexts, see Barnette & McLean, 2005; Klockars & Hancock, 1994; Shaffer, Kowalchuk, & Keselman, 2013).

Methodology

Monte Carlo simulations were conducted in R (R Core Team, 2013) of two-group designs with 50 subjects per group. Three numbers of multivariate normal outcome variables were used: $m = 2$, $m = 5$, and $m = 10$. 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

At each value of *m*, each of the four procedures had a maximum FWER less than .050, but the PFER could differ notably from the FWER when outcome variables were correlated. For example, Figure 1B 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 and is insensitive to correlation.

Note that in Figures 1B and 1C, 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, Kowalchuk, & Keselman, 2013) showed that the PFER can be substantially inflated in multigroup designs even when the FWER is controlled. This article 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.

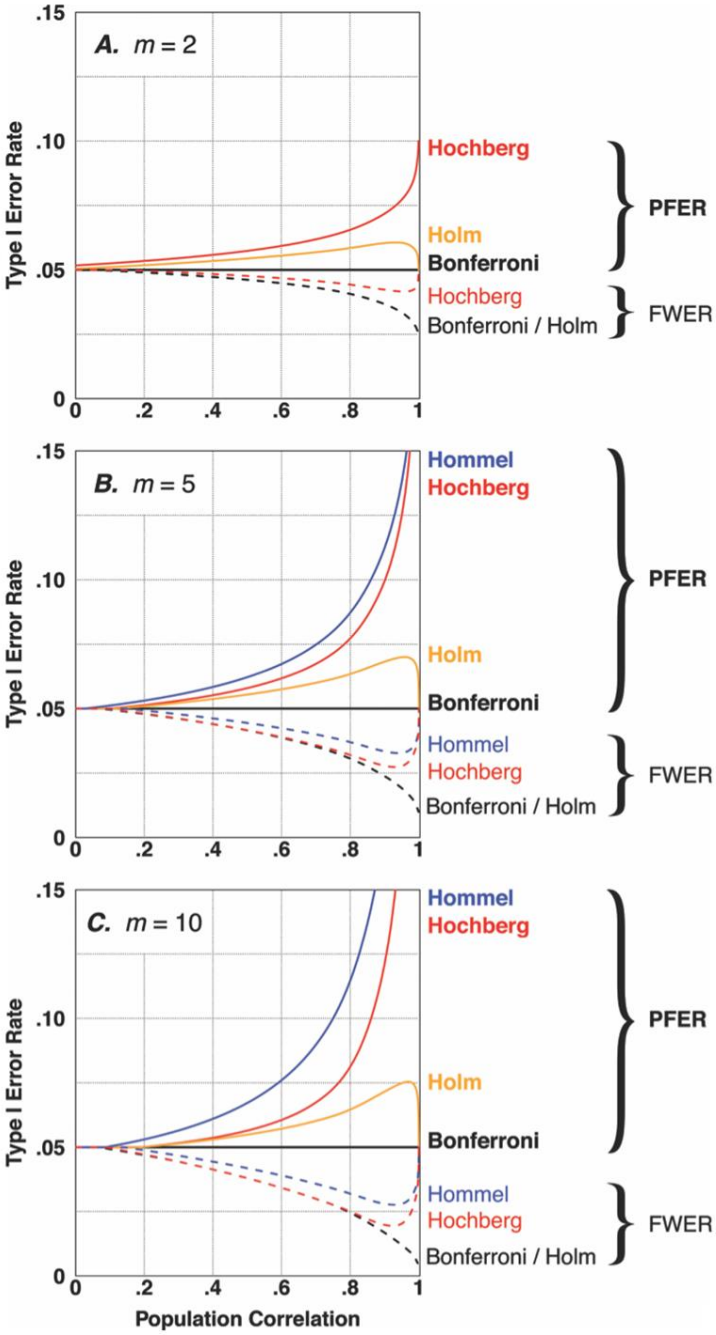

Figure 1. Per-family and familywise Type I error rates for the Bonferroni, Holm, Hochberg, and Hommel procedures in a two-group design with *m* outcome variables (α = .05, all null hypotheses true). Note that Hommel is equivalent to Hochberg for m = 2.

Implications for research practice

This article 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 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 the toleration of false findings.

Controlling the FWER may be sufficient when, given one Type I error, additional Type I errors are not costly, or perhaps when dependency among the tests is known to be sufficiently low that FWER and PFER are only negligibly different. In such situations, a method more powerful than the Bonferroni procedure may be used, such as the Holm procedure (if confidence intervals are required), the Hochberg or Hommel procedure (if no confidence intervals are required), or a context specific method appropriate for the given situation (see Dmitrienko et al., 2010 for an extensive list). An important caveat is that the Hochberg and Hommel procedures do not necessarily control the FWER for one sided tests that can be negatively correlated (see Samuel-Cahn, 1996), whereas the Bonferroni and Holm methods do not have this limitation.

Implications for applied statistics pedagogy

If the PFER is to be addressed more in practice, then it must also be addressed more in pedagogy. Therefore, this article recommends that professors and textbook authors 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 no longer sufficient to simply refer to "the Type I error rate."

Limitations

This study did not examine every Type I error rate that has been defined. For example, the comparisonwise Type I error rate (Tukey, 1953) is the probability of Type I error for a single hypothesis irrespective of the number of hypotheses in the family. Thus, controlling Type I error at the comparisonwise level effectively means disregarding Type I error inflation altogether and simply conducting each

hypothesis test at the unadjusted alpha level. Another error rate that has been proposed is the false discovery rate (Benjamini & Hochberg, 1995), which is, loosely speaking, the expected proportion of significant hypothesis tests in the family that are Type I errors (except when all null hypotheses are true, in which case the false discovery rate is equivalent to the FWER). Both the comparisonwise Type I error rate and the false discovery rate are more liberal than the FWER and thus beyond the scope of this article, but there are contexts in which these error rates may be appropriate.

It should also be acknowledged that the simulations examined neither a variety of alpha levels, nor an exhaustive variety of multiple comparisons procedures, nor an exhaustive variety of parameter combinations. However, to do so would have made exceedingly long and complex an article that required only a finite number of examples to support its conclusion that the PFER can be relevant. Future articles may examine in detail issues such as which Type I error rates are more relevant in particular contexts.

References

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. Available at http://digitalcommons.wayne.edu/jmasm/vol4/iss2/10/

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 (Methodological), 57*(1), 289-300. Retrieved from http://www.jstor.org/stable/2346101

Bird, K. D., & Hadzi-Pavlovic, D. (2013). Controlling the maximum familywise Type I error rate in analyses of multivariate experiments. *Psychological Methods*. Advance online publication. 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

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. (2010). Multiple testing methodology. In A.

Dmitrienko, A. C. Tamhane, & F. Bretz (Eds.), *Multiple testing problems in pharmaceutical statistics* (pp. 35-98). Boca Raton, FL: Chapman & Hall.

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

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:10.3233/NRE-130893

Goodwin, C. J. (2010). *Research in psychology: Methods and design* (6th ed.). Hoboken, NJ: John Wiley & Sons.

Gordon, A., Glazko, G., Qiu, X. & 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:10.1214/07-AOAS102

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

Harris, R. J. (2001). *A primer of multivariate statistics* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

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. *Scandanavian Journal of Statistics, 6*, 65-70.

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. (2014). *Fundamental statistics for the behavioral sciences* (8th ed.). Belmont, CA: Wadsworth.

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

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

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, England: Sage.

Miller, R. G., Jr. (1966). *Simultaneous statistical inference*. New York, NY: McGraw-Hill.

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

R Core Team. (2013). R (Version 3.0.2) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika*, *77*(3), 663-665. doi:10.1093/biomet/77.3.663

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

Samuel-Cahn, E. (1996). Is the Simes improved Bonferroni procedure conservative? *Biometrika*, *83*(4), 928-933. doi:10.1093/biomet/83.4.928

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., 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

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, *73*(3), 751-754. doi:10.1093/biomet/73.3.751

Sirkin, R. M. (2006). *Statistics for the social sciences* (2nd ed.). Thousand Oaks, CA: Sage.

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

Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.

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.

Wetcher-Hendricks, D. (2011). *Analyzing quantitative data*. Hoboken, NJ: John Wiley & Sons.