# Neglect of Multiplicity in Hypothesis Testing of Correlation Matrices

Burt Holland

Department of Statistics

Temple University

Philadelphia, PA 19122

*e-mail*: bholland@temple.edu

examining overestimation and underestimation is commonplace in the literature, it may lead to misinterpretation. For example, a positive coefficient linking an information source to the difference score may suggest that as applicants increasingly rely on a recruitment source, they (1) move from underestimation to agreement or (2) move from agreement to overestimation. To clarify the difference score results, we conducted a second analysis in which we regressed applicants' culture beliefs on the information sources. We used the resulting slopes and intercepts to determine whether our findings indicated overestimation or underestimation and to test whether applicants' and executives' culture beliefs were significantly different at low versus high levels of information source use.

***Accuracy.*** We also performed two types of analyses to examine the accuracy of applicants' culture beliefs. First, for each value, we computed the absolute difference between executives' beliefs and applicants' beliefs, and we regressed these absolute difference scores on the five information sources. Despite the prevalence of this approach, it also can obscure the true nature of the relationships under study, leading to misinterpretation. For instance, a negative relationship between an information source and the absolute difference score would imply increased accuracy. Fundamentally, increased accuracy means that, as use of an information source increases, the culture belief scores of applicants whose scores fell below executives' scores increase, and the scores of applicants that were above the executives' decrease. An absolute difference score cannot reveal whether both of these effects, in fact, exist. Therefore, we also used a regression procedure in which a dummy variable distinguished two subgroups of applicants, one whose culture belief scores fell below those of the executives, and another whose culture belief scores were above those of the executives (Edwards, 1995). The dummy variable was used as a moderator to determine whether the slopes differed for the two groups of applicants in a way that is consistent with the meaning of accuracy, so that the scores of applicants whose beliefs fell below executives' increased and the scores of applicants whose beliefs were above the executives' decreased. If these moderator analyses indicated that slopes did not differ for the two subgroups of applicants, we combined the subgroups to estimate a single relationship between information sources and applicants' culture beliefs. For this combined group, accuracy was represented by the distance between the executives' beliefs and the regression line relating an information source to applicants' beliefs. If this distance decreased significantly as use of the information source increased, then evidence for accuracy was obtained. Note that a decrease in distance is evidenced by either: (1) a negative slope, with an intercept above executives' beliefs, or (2) a positive slope, with an intercept below executives' beliefs.

## RESULTS

Table 1 lists the means, standard deviations, and correlations among the variables. Table 2 shows the results from using both the algebraic difference score approach and the regression approach to predict whether applicants' beliefs represented overestimation or underestimation of Company X's culture relative to the executives' beliefs. Note that the coefficients between the difference score and regression approaches are identical, and that only the intercepts are different. Finally, Table 3 shows the accuracy results from using either the absolute difference score approach or the moderated regression approach. In Tables 2 and 3, we have highlighted the particular results that correspond to each of our hypotheses.

***Company information.*** Hypothesis 1a suggests that applicants would overestimate the degree to which Company X valued risk taking when they relied on company information sources. As shown

**TABLE 1**
**Means, Standard Deviations, and Correlations between Variables[a]**

| Variable | Mean | s.d. | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|---|
| 1. Applicants' risk perceptions | 4.65 | 1.13 | | | | | | |
| 2. Applicants' rules perceptions | 4.35 | 1.11 | −.22 | | | | | |
| 3. Applicants' results perceptions | 5.65 | 1.00 | .37 | −.10 | | | | |
| 4. Company information | 3.43 | 1.28 | .26 | −.09 | .30 | | | |
| 5. Product ads/products | 4.88 | 1.44 | .20 | .01 | .21 | .31 | | |
| 6. Company experience | 1.79 | 1.96 | −.05 | .14 | −.03 | .09 | .09 | |
| 7. Word of mouth | 3.01 | 1.43 | .05 | .13 | .13 | .45 | .32 | .21 |

[a] $n = 240$; correlations greater than .12 are significant at the .05 level, under two-tailed tests.

**TABLE 1**
**Descriptive Statistics and Correlations for Study Variables[a]**

| Variable[b] | Mean | s.d. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. $Ability_{PM}$ | 3.45 | 0.82 | (.91) | | | | | | | | | | | |
| 2. $Benevolence_{PM}$ | 3.14 | 0.93 | .65** | (.92) | | | | | | | | | | |
| 3. $Integrity_{PM}$ | 3.23 | 0.85 | .73** | .80** | (.89) | | | | | | | | | |
| 4. $Ability_{TMT}$ | 2.94 | 0.70 | .27** | .15* | .14* | (.89) | | | | | | | | |
| 5. $Benevolence_{TMT}$ | 2.59 | 0.75 | .23** | .35** | .23** | .55** | (.87) | | | | | | | |
| 6. $Integrity_{TMT}$ | 2.79 | 0.63 | .20** | .27** | .25** | .66** | .78** | (.85) | | | | | | |
| 7. $Trust_{PM}$ | 3.21 | 0.77 | .74** | .72** | .76** | .13* | .20** | .19** | (.81) | | | | | |
| 8. $Trust_{TMT}$ | 2.72 | 0.63 | .22** | .17** | .15* | .62** | .66** | .71** | .26** | (.72) | | | | |
| 9. Ability to focus | 2.66 | 0.88 | .29** | .42** | .40** | .23** | .31** | .32** | .35** | .24** | (.77) | | | |
| 10. In-role performance | 4.05 | 0.61 | .03 | .10 | .11 | −.08 | −.05 | .02 | .03 | −.05 | .09 | (.91) | | |
| 11. OCBI | 2.94 | 0.64 | .12 | .20** | .16* | −.10 | .05 | .01 | .19** | −.04 | .15* | .50** | (.82) | |
| 12. OCBO | 3.82 | 0.62 | .19** | .22** | .23** | .02 | .10 | .17** | .16* | .12* | .19** | .62** | .42** | (.67) |

[a] $n = 247$. Alpha coefficients are on the diagonal in parentheses.

[b] The subscript "PM" indicates that an employee's plant manager was the referent for the designated measure. The subscript "TMT" indicates that the referent was the studied firm's top management team. "OCBI" is organizational citizenship behavior directed toward individuals, and "OCBO" is organizational behavior directed toward one's organization.

* $p \leq .05$
** $p \leq .01$

**TABLE 2**
**Results of Structural Nested Model Comparisons[a]**

| Model | $\chi^2$ | *df* | CFI | RMSEA | SRMSR | $\Delta\chi^2$ *(df)*[b] |
|---|---|---|---|---|---|---|
| Hypothesized model | 3,668.92** | 2,175 | .96 | .05 | .06 | |
| Hypothesized model without correlated residuals between in-role performance, OCBI and OCBO | 3,866.78** | 2,178 | .95 | .05 | .09 | 197.86* (3) |
| Hypothesized model less the six direct effects relating $trust_{PM}$ and $trust_{TMT}$ to in-role performance, OCBI, and OCBO | 3,680.95** | 2,181 | .96 | .05 | .07 | 12.03 (6) |
| Hypothesized model less the $trust_{PM} \rightarrow$ in-role performance direct effect | 3,668.92** | 2,176 | .96 | .05 | .06 | 0.00 (1) |
| Hypothesized model less the $trust_{TMT} \rightarrow$ in-role performance direct effect | 3,670.31** | 2,176 | .96 | .05 | .06 | 1.39 (1) |
| Hypothesized model less the $trust_{PM} \rightarrow$ OCBI direct effect | 3,672.13** | 2,176 | .96 | .05 | .07 | 3.23 (1) |
| Hypothesized model less the $trust_{TMT} \rightarrow$ OCBI direct effect | 3,670.78** | 2,176 | .96 | .05 | .06 | 1.86 (1) |
| Hypothesized model less the $trust_{PM} \rightarrow$ OCBO direct effect | 3,671.10** | 2,176 | .96 | .05 | .07 | 2.18 (1) |
| Hypothesized model less the $trust_{TMT} \rightarrow$ OCBO direct effect | 3,669.06** | 2,176 | .96 | .05 | .06 | 0.14 (1) |
| Hypothesized model less the $trust_{PM} \rightarrow$ in-role performance and $trust_{TMT} \rightarrow$ in-role, direct effects | 3,670.32** | 2,177 | .96 | .05 | .06 | 1.40 (2) |
| Hypothesized model less the $trust_{PM} \rightarrow$ OCBI and $trust_{TMT} \rightarrow$ OCBI direct effects | 3,673.99** | 2,177 | .96 | .05 | .07 | 5.07 (2) |
| Hypothesized model less the $trust_{PM} \rightarrow$ OCBO and $trust_{TMT} \rightarrow$ OCBO direct effects | 3,671.24** | 2,177 | .96 | .05 | .07 | 2.32 (2) |

[a] CFI is the comparative fit index; RMSEA is the root-mean-square error of approximation; SRMSR is the standardized root-mean-square residual.

[b] The subscript "PM" indicates that an employee's plant manager was the referent for the designated measure. The subscript "TMT" indicates that the referent was the studied firm's top management team. "OCBI" is organizational citizenship behavior directed toward individuals, and "OCBO" is organizational citizenship behavior directed toward one's organization.

* $p \leq .05$
** $p \leq .01$

The Academy of Management

# Journal

Volume 43
Number 6

December 2000



*Continued on back cover*

Academy Of Management

- Many papers appearing in social science journals contain sample correlation matrices accompanied by individual tests that the corresponding correlation parameters are zero.

- For example, eight of the ten papers in the October 2005 issue of the *Academy of Management Journal* contain sample correlation matrices accompanied by stars or daggers indicating which correlations differ significantly from zero.

- These tests ignore the issue of multiplicity in a related family of inferences.

- Ignoring FWE control in the context of correlation leads to spurious determinations of linear relationships among study variables.

- The 2006 Journal Impact Factor of *Academy of Management Journal* was 3.353. By comparison, the Impact Factor of *JASA* was 2.171.

- In most Management articles when correlation matrices are included with indications of significance, it is for descriptive purposes only. In some Management articles with correlation matrices, such significance tests relate directly to the article's research hypotheses.

- Should social scientists who present such matrices be required to give $p$-values adjusted for multiplicity?

An exception to my statement about social science literature is Educational and Psychological research. For example, Olejnik, Li, Supattathnum & Huberty (1997) explicitly discussed the need to control for multiplicity in the context of correlation matrices.

*Journal of Educational and Behavioral Statistics*
*Winter 1997, Vol. 22, No. 4, pp. 389–406*

# Multiple Testing and Statistical Power With Modified Bonferroni Procedures

**Stephen Olejnik**
*University of Georgia*

**Jianmin Li**
*TRW*

**Suchada Supattathum**
*King Mongleut Institute of Technology*

**Carl J Huberty**
*University of Georgia*

Keywords: *Bonferroni, familywise error rate, multiple testing*

*The difference in statistical power between the original Bonferroni and five modified Bonferroni procedures that control the overall Type I error rate is examined in the context of a correlation matrix where multiple null hypotheses,* $H_0$: $\rho_{ij} = 0$ *for all* $i \neq j$, *are tested. Using 50 real correlation matrices reported in educational and psychological journals, a difference in the number of hypotheses rejected of less than 4% was observed among the procedures. When simulated data were used, very small differences were found among the six procedures in detecting at least one true relationship, but in detecting all true relationships the power of the modified Bonferroni procedures exceeded that of the original Bonferroni procedure by at least .18 and by as much as .55 when all null hypotheses were false. The power difference decreased as the number of true relationships decreased. Power differences obtained for the average power were of a much smaller magnitude but still favored the modified Bonferroni procedures. For the five modified Bonferroni procedures, power differences less than .05 were typically observed; the Holm procedure had the lowest power, and the Rom procedure had the highest.*

There have been several discussions (Keppel, 1991; Maxwell & Delaney, 1990; Ryan, 1959; Toothaker, 1991) on the issue of controlling the overall Type I error rate (the probability of at least one Type I error in a set of hypotheses) in situations where multiple tests are conducted simultaneously (e.g., comparisons of means, factorial ANOVA, factorial ANOVA with multiple outcome variables, multiple chi-squared tests). To provide this control, the significance level used to test an individual hypothesis is modified (adjusted) to take into consideration the number of hypotheses tested. The simplest and perhaps the best known procedure is to divide the acceptable overall risk of a Type I error by the number of hypotheses tested. This approach is based on the Bonferroni inequality:

In the Ethical Guidelines for Statistical Practice at the website of the American Statistical Association, Section II.A.8 reads:

Recognize that any frequentist statistical test has a random chance of indicating significance when it is not really present. Running multiple tests on the same data set at the same stage of an analysis increases the chance of obtaining at least one invalid result. Selecting the one "significant" result from a multiplicity of parallel tests poses a grave risk of an incorrect conclusion. Failure to disclose the full extent of tests and their results in such a case would be highly misleading.

This is serious criticism of the article formatting requirements of *Academy of Management Journal* and other journals in the Social Sciences.

- Most multiple comparison procedures that control FWE require that the joint distribution of the family of test statistics satisfy the very general assumption of subset pivotality.

- Subset pivotality: the joint distribution of the family p-values is not dependent on what subset of the family null hypotheses is true.

- This assumption is not met by the joint distribution of the sample correlations comprising a correlation matrix. As a result, very few multiple testing procedures have been proven to control FWE when simultaneously testing correlations.

- **Notation**: Assume there is a family of $k$ related inferences. In the context of an $m \times m$ correlation matrix, $k$ is the number of distinct correlations in such a matrix, $m(m-1)/2$. The population coefficient of correlation between variables $i$ and $j$ is denoted $\rho_{ij}$. Conventionally, researchers simultaneously consider the family of $k$ related hypothesis tests

  $$H_0\colon \rho_{ij} = 0 \quad \text{vs} \quad H_1\colon \rho_{ij} \neq 0$$

  for $1 \leq j < i \leq m$. Let the ordered ordinary raw $p$-values for the $k$ tests be $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(k)}$ with corresponding hypotheses $H_1, H_2, \ldots H_k$.

- I discuss three methods for controlling familywise Type I error that are applicable to simultaneously testing all correlations in a correlation matrix. One of these methods is extremely easy to implement.

The three methods are:

- Holm (1979)
- Westfall & Young (1993)
- Romano & Wolf (2005)

The Holm procedure, which I abbreviate as Holm, is very easily to conduct; it can be managed by hand for small families of tests. The two newer procedures use computer intensive bootstrap methodology, (Efron & Tibshirani (1993). I provide descriptive algorithms for implementing all three. Unlike the bootstrap procedures, Holm ignores the dependence between the correlations and requires only knowledge of the raw (unadjusted) univariate $p$-values and family size $k$. Holm constitutes a modest improvement of the familiar and widely used Bonferroni procedure.

**The Holm procedure**: Westfall & Young (1993) provide explicit formulas for the adjusted $p$-values, $p_{Hi}$ of the easily conducted procedure of Holm (1979):

$$\begin{aligned} p_{H1} &= kp_{(1)} \\ p_{H2} &= \max(p_{H1}, (k-1)p_{(2)}) \\ \vdots & \quad \vdots \\ p_{Hi} &= \max(p_H(i-1), (k-i+1)p_{(i)}) \\ \vdots & \quad \vdots \\ p_{Hk} &= \max(p_H(k-1), p_{(k)}) \end{aligned}$$

The "max" portion of these formulas ensures that the adjusted $p$-values are monotonically non-increasing. If these formulas lead to any adjusted $p$-values in excess of 1, these adjusted $p$-values should be truncated to 1.

Beginning from a data set consisting of a column of the $k$ unadjusted $p$-values, Westfall, Tobias, Rom, Wolfinger & Hochberg (1999) demonstrate how to produce a column of the $k$ Holm adjusted $p$-values using just 9 lines of SAS code. SAS users can mimic this Westfall et al (1999) example to easily calculate the Holm adjusted $p$-values for their correlation matrix.

The paper contains algorithms for implementing the other two procedures, Westfall & Young and Romano & Wolf.

In the context of testing whether elements of a correlation matrix are zero, when will Holm suffice and when should one make the effort to instead use one of the other two? When is the Romano & Wolf procedure preferred to the Westfall & Young procedure?

There is no statistical theory that strictly ranks the procedures according to a MCP power concept. However, the Westfall & Young algorithm is based on the assumption that the tests are independent. The Westfall & Young procedure is more conservative and less powerful than the Romano & Wolf procedure for families with greater departure from independence.

- My recommendations are based on examinations of numerous simulations of all procedures covering a wide variety of contrived and real correlation matrices of various dimensions, and involving sample sizes from small to intermediate.

- My simulation program, which was written in *R* software, matched to 3 decimal places the bootstrap $p$-values for the analysis of a real data set summarized in Table 4 of Romano & Wolf (2005). Romano & Wolf use this example to suggest that their procedure tends to be more powerful than Westfall & Young's, but they do not prove uniformly greater power.

- While my simulations demonstrate that the Romano & Wolf procedure occasionally rejects hypotheses not rejected by the Westfall & Young procedure, I have seen occasions where these two procedures agree on rejections and occasions where, for some of the correlations, Westfall & Young adjusted $p$-values fall below the corresponding ones calculated using the Romano & Wolf procedure.

In most instances, data analysis by social scientists involving correlation matrices no larger than $15 \times 15$ can use Holm and avoid the burden of a heavy duty statistical analysis.

There is one situation when one of the bootstrap procedures, preferably Romano & Wolf's, should be considered for use because it tends to declare more correlations significantly different from zero than does Holm. This occurs when the $m$ variables are moderately or severely collinear, as when there is a near redundancy involving the chosen variables. Collinearity impacts the correlation matrix by giving it a high condition index, defined as the ratio of largest to smallest eigenvalue, a number calculated from the matrix itself. The relevance of condition index as an indication of which procedure to use is illustrated with examples in the paper.

## Example with a moderate-sized correlation matrix

Table 1 in Mayer & Gavin (2005) is a correlation matrix involving 12 variables having condition index 31.72, indicating moderate collinearity among these variables. The sample size was 247. Using conventional tests on the 66 correlations based on significance level $\alpha = .05$, Mayer & Gavin determined that 50 of these correlations differ significantly from zero.

I did not have access to the data set, only to the correlation matrix. I used the `R` function `mvrnorm` to simulate 247 multivariate normal obserations having the same correlation matrix. After awhile I produced a correlation matrix that was reasonally close. From the 247 observations having this correlation structure, I performed $B = 5,000$ bootstrap samples. I repeated this exercise many times.

Comment: I don't know if Mayer & Gavin's data is MVN. However, it is very likely that they assumed bivariate normal distributions to calculate their unadjusted $p$-values.

Controlling FWE at .05, both the Holm and the Westfall & Young procedures found that just 41 correlations differ significantly from zero while the Romano & Wolf procedure detected 42 correlations differing significantly from zero. Comparing the sizes of $p$-values, most Westfall & Young adjusted $p$-values were smaller than the corresponding Holm adjusted $p$-values, and the 66 Romano & Wolf adjusted $p$-values were uniformly smaller than the Westfall & Young adjusted $p$-values.

## Example with a smaller correlation matrix

- The correlation matrix in Table 1 in Cable, Aiman-Smith, Mulvey & Edwards (2000) is $7 \times 7$, condition index 4.85, calculated from a sample size of $n = 240$.
- For each of the 21 variable pairs, I provide in this presentation's Table 1 the correlations given in Cable et al (2000) and, for a correlation matrix very similar to the one in Cable et al (2000), the adjusted $p$-values for each of the three FWE controlling methods. Since the original correlations were stated with 2 digit accuracy, at most 2 digit accuracy may be claimed for my presented adjusted $p$-values.
- Cable et al (2000) stated that the 13 correlations having absolute value exceeding .12 are significant at the .05 level. Examining columns 3–5 of Table 1, we note that the procedures of Holm, Westfall & Young, and Romano & Wolf all have 9 adjusted $p$-values at most .05 and therefore all reject (the same) 9 hypotheses when controlling FWE at .05. Considering the 21 rows of adjusted $p$-values, we see that the Romano & Wolf procedure usually has the lowest $p$-value and that the Holm procedure never does.

## Summary and Conclusions

- The easily implemented Holm procedure will usually suffice. The larger the condition index of the correlation matrix, that is, the closer is one or more relationships among study variables to exactly linear, the more likely that the Romano & Wolf procedure will reject more hypotheses of zero correlation than the other two procedures, and the more likely that the Westfall & Young procedure will reject hypotheses not rejected by Holm.
- The study's sample size used does not impact the preceding statement. However, larger sample sizes lead to lowered adjusted $p$-values and potentially more rejections for all three procedures.
- I would appreciate additional suggestions for selling social scientists the idea of controlling for multiplicy in correlation matrices.
- I would appreciate suggestions for a venue to publish this idea.

## SELECTED REFERENCES


1. Cable, D.M., Aiman-Smith, L., Mulvey, P.W. and Edwards, J.R., (2000). The Sources and Accuracy of Job Applicants' Beliefs About Organizational Culture. *Academy of Management Journal*, **43**, 1076–1085.

2. Efron, B. and Tibshirani, R. (1993). An Introduction to the Bootstrap, London: Chapman and Hall.

3. Holm, S., (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

4. Mayer, R.C. and Gavin, M.B., (2005). Trust in Management and Performance: Who Minds the Shop while the Employees Watch the Boss? *Academy of Management Journal*, **48**, 874–888.

5. Olejnik, S., Li, J., Supattathnum, S, and Huberty, C.J. (1997). Multiple Testing and Statistical Power with Modified Bonferroni Procedures. *Journal of Educational and Behavioral Statistics,* **22**, 389–406.

6. Romano, J.P. and Wolf, M. (2005). Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing. *J. American Statist. Assoc.* **100**, 94–108.

7. Westfall, P. H. and Young, S. S., (1993). Resampling-Based Multiple Testing, New York: Wiley.

8. Westfall, P. H., Tobias, R. D., Rom, D., Wolfinger, R.D., and Hochberg, Y., (1999). Multiple Com-

parisons and Multiple Tests Using the SAS System, Cary, NC: SAS Institute, Inc.

Table 1: Correlations and $p$-values for the $7 \times 7$ correlation matrix in Cable et al (2000). Notation: $p_{Hi}, p_{Wi}, p_{Ri}$ are the adjusted $p$-values using the methods of Holm (1979), Westfall & Young (1993), and Romano & Wolf (2005) respectively.

| variable pair | correlation | $p_{Hi}$ | $p_{Wi}$ | $p_{Ri}$ |
|---|---|---|---|---|
| (V1,V2) | -.22 | .0019 | .0018 | .0016 |
| (V1,V3) | .37 | .0000 | .0000 | .0000 |
| (V1,V4) | .26 | .0000 | .0000 | .0000 |
| (V1,V5) | .20 | .1957 | .1802 | .1850 |
| (V1,V6) | -.05 | 1.0000 | .7952 | .7792 |
| (V1,V7) | .05 | 1.0000 | .9782 | .8654 |
| (V2,V3) | -.10 | .0446 | .0494 | .0442 |
| (V2,V4) | -.09 | .3699 | .3196 | .3058 |
| (V2,V5) | .01 | 1.0000 | .9782 | .9770 |
| (V2,V6) | .14 | .2170 | .1964 | .1976 |
| (V2,V7) | .13 | .9070 | .6132 | .5984 |
| (V3,V4) | .30 | .0002 | .0000 | .0000 |
| (V3,V5) | .21 | .0060 | .0052 | .0052 |
| (V3,V6) | -.03 | 1.0000 | .8252 | .7684 |
| (V3,V7) | .13 | 1.0000 | .7896 | .7724 |
| (V4,V5) | .31 | .0002 | .0000 | .0000 |
| (V4,V6) | .09 | 1.0000 | .9724 | .9738 |
| (V4,V7) | .45 | .0000 | .0000 | .0000 |
| (V5,V6) | .09 | 1.0000 | .8252 | .8218 |
| (V5,V7) | .32 | .0000 | .0000 | .0000 |
| (V6,V7) | .21 | .0955 | .0922 | .0938 |