# Multiplicity control in Structural Equation Modeling:

# Incorporating Parameter Dependencies

Carrie Smith & Robert A. Cribbie

Quantitative Methods Program

Department of Psychology

York University

Please send comments regarding this manuscript to: Carrie Smith, Quantitative Methods Program, Department of Psychology, York University, Toronto, ON, M3J 1P3, smithce@yorku.ca

Abstract

When structural equation modeling (SEM) analyses are conducted, significance tests for all important model relationships (parameters including factor loadings, covariances, etc.) are typically conducted at a specified nominal Type I error rate (α). Despite the fact that many significance tests are often conducted in SEM, rarely is multiplicity control applied. Cribbie (2000, 2007) demonstrated that without some form of adjustment, the familywise Type I error rate can become severely inflated. Cribbie also confirmed that the popular Bonferroni method was overly conservative due to the correlations among the parameters in the model. The purpose of this study was to compare the Type I error rates and per-parameter power of traditional multiplicity strategies with those of adjusted Bonferroni procedures that incorporate not only the number of tests in a family, but also the degree of correlation between parameters. The adjusted Bonferroni procedures were found to produce per-parameter power rates higher than the original Bonferroni procedure without inflating the familywise error rate.

*Keywords*: Multiplicity control, Type I error rates

Significance Testing in Structural Equation Modeling:

Incorporating Parameter Dependencies into Multiplicity Controlling Procedures

Structural equation modeling (SEM) is capable of assessing models in which complex multivariate interrelationships are hypothesized, and is therefore well suited to address research questions that are typical within psychology, education, business and other related fields. SEM has steadily gained popularity and a large body of research has emerged around fit indices, estimation methods, and assumption violations. However, relatively little attention has been paid to the issue of inflated Type I error rates when the statistical significance of multiple parameters is evaluated within a model.

When a large number of pairwise comparisons are conducted in a mean difference analysis, some form of error control is generally considered necessary and is almost universally applied (e.g., post hoc tests for a one-way ANOVA). It is commonly understood that when *k* tests, where *k* represents the number of pairwise tests conducted (not the number of means being compared), are performed at a specific nominal Type I error rate ($\alpha_{\text{per test}}$), that the probability of committing one or more Type I errors within the set or 'family' of tests ($\alpha_{\text{familywise}}$) increases with the number of comparisons.

If all of the tests conducted are independent, then the overall error rate becomes $\alpha_{\text{familywise}} = 1 - (1 - \alpha_{\text{per test}})^k$, which is roughly equal to $k\alpha_{\text{per test}}$. Note that independent tests represent the worst-case in terms of the expected number of Type I errors. If the tests were *completely* dependent, then $\alpha_{\text{familywise}}$ would equal $\alpha_{\text{per test}}$, since if one test was in error, so are all the rest. Correlated tests ($0 < |\rho| < 1$) therefore fall between these two possibilities, and if no multiplicity control is imposed and all correlated tests are

conducted at $\alpha_{\text{per test}}$, $\alpha_{\text{familywise}}$ will become inflated. This situation is generally considered unacceptable and familywise error control is usually imposed by adjusting $\alpha_{\text{per test}}$ for each of the pairwise comparisons, holding $\alpha_{\text{familywise}}$ at an acceptable threshold, albeit at the expense of statistical power.

In contrast, consider a typical structural model. After an adequate overall model fit has been obtained, significance tests for all important model relationships (e.g., directional and non-directional structural model parameters) are conducted at a specified $\alpha$, such as .05. Despite the fact that many significance tests are often conducted in SEM, rarely is multiplicity control applied as it was in the preceding example. A prominent reason for this is the claim that because parameters in SEM are likely to be correlated, traditional methods of Type I error control are too conservative (Mulaik, 2004; Owen, 2004; Ronis, 2002). Cribbie (2000, 2007) demonstrated that without some form of adjustment, the familywise Type I error rate can become severely inflated. Cribbie also confirmed that the popular Bonferroni method was overly conservative due to the correlations among the parameters in the model, and recommended more powerful familywise and false discovery rate controlling methods.

One important limitation of the previous studies by Cribbie (2000, 2007) is that there was no correction for the degree of dependency among the parameters. Therefore, the purpose of this study is to investigate potential adjustments to the traditional Bonferroni procedure that incorporate information about the degree of relationship among the parameters in the model. A simulation study will be conducted to compare the familywise error rates and per-parameter power of the adjusted Bonferroni procedures, in

contrast with no multiplicity control, and multiplicity control imposed via the traditional Bonferroni method.

*Adjusted Bonferroni Procedures*

The original Bonferroni procedure is a strategy for controlling $\alpha_{familywise}$ at $\alpha$ for a set of null hypothesis tests, where $\alpha_{per\ test}$ is set equal to $\alpha_{familywise} / k$. Although the Bonferroni procedure is effective at providing strict $\alpha_{familywise}$ control, the probability of Type II errors can become extremely high, especially when a large number of tests are conducted and/or the tests are highly related. Thus, the original Bonferroni procedure is not an ideal multiple testing procedure for SEM where there are often numerous parameters to be investigated and the correlations among the parameters are often high.

In this study we explore two adjusted Bonferroni procedures that incorporate the degree of interrelationship among the parameters in the model when establishing an appropriate adjusted $\alpha_{per\ test}$. Both adjusted Bonferroni methods were derived under the theory that if *k* null hypothesis tests are related then the optimal $\alpha_{per\ test}$ is not $\alpha_{familywise} / k$, but instead a less stringent correction that incorporates the degree of interrelatedness among the parameters. More specifically, the two functions divided $\alpha_{familywise}$ across the *k* tests similar to the Bonferroni procedure, however the severity of the adjustment was weakened with an increasing value of the average absolute correlation of parameter *j* with other parameters in the model ($\overline{|r_j|}$). The average absolute correlation ($\overline{|r_j|}$) of parameter *j* was chosen as an index of the extent to which the given parameter covaries with other parameters in the model.

Adjusted Bonferroni 1 (AB1):

$$\alpha_{per\ test} = \frac{\alpha_{familywise}}{k - (k - 1)\sqrt{|r_j|}}.$$

Adjusted Bonferroni 2 (AB2):

$$\alpha_{per\ test} = \frac{\alpha_{familywise}}{k^{1-\sqrt{|r_j|}}}.$$

Both of the derived functions respect two important properties. First, if the correlation between tests is 0 (completely independent), the adjustments are equivalent to $\alpha_{familywise}$ divided by *k* (the number of tests), which is equivalent to performing a traditional Bonferroni adjustment. If the correlation between tests is equal to 1, indicating that they are perfectly correlated, the adjustment is equivalent to 1, indicating no adjustment for multiplicity.

## Method

A Monte Carlo study was conducted to compare the Type I error rates and per parameter power when no multiplicity control (NC), Bonferroni control, or adjusted Bonferroni control (AB1, AB2) was imposed for evaluating the statistical significance of multiple parameters in the structural portion of a latent variable model.

Simulations were conducted for two models. Model A (Figure 1) had 237 degrees of freedom with 15 hypothesis tests in the structural model of which 4 were set to be null (the dashed lines in Figure 1). These paths were fixed to zero in the population covariance matrix. Further, the latent variable variances were set to 1, factor loadings to 0.8 and error variances to 0.36.

Model B (Figure 2) had 565 degrees of freedom with 29 hypothesis tests in the structural model of which 6 were null. Type I error control was applied over all

parameters in the structural model, thus the family of selected tests over which Type I error control was of size $k = 15$ for Model A and $k = 29$ for Model B. For the simulations model specification was achieved by fixing one factor loading per latent variable to one.

Two primary variables were manipulated in this study: 1) the correlations among the parameters; and 2) sample size. The correlations between parameters were manipulated indirectly by varying the magnitude of the correlation between the latent variables in the population from which the samples were generated ($r_{latents}$ = 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45). Sample sizes utilized were $N = 100$ and $N = 400$.

Type I error control was evaluated with respect to familywise error rates (i.e., the probability of declaring at least one null parameter statistically significant). Per-parameter power was evaluated as the average proportion of nonzero parameters declared statistically significant.

The simulation was conducted using R (R Development Core Team, 2010) employing the 'sem' package (Fox, 2006). One thousand simulations were conducted for each condition using a nominal significance level of $\alpha = .05$. Two fit indices were utilized to verify that the fit of the models was acceptable, the Comparative Fit Index (CFI; Bentler, 1990) and the Root Mean Squared Error of Approximation (RMSEA; Steiger & Lind, 1980).

## Results

The pattern of results for the fit indices, familywise error rates and power were very similar across Models A and B and therefore only the results for Model B (with more null paths and therefore a greater probability of inflated familywise error rates) are discussed.

### *Fit Indices*

The RMSEA and CFI indicated an excellent fit of the models to the data, with CFI > .97 and RMSEA < .03 across all conditions investigated.

*Familywise Type I Error Rates*

Familywise error rates for N = 100 and 400, across the levels of parameter interrelatedness, are displayed in Figure 3. As expected, familywise Type I error rates far exceeded the per-parameter alpha of .05 when no multiplicity control (NC) was imposed, reaching values in excess of 0.25. The Bonferroni procedure was uniformly conservative, with familywise error rates hovering around .01. AB2 showed a very slight increase in familywise error rates over the Bonferroni procedure, while error rates for AB1 ranged from 2 to 3%. Both were below the nominal α-level of .05.

*Per-Parameter Power*

The average per-parameter power for Model B with $N$ = 100 and 400, for a subset of the levels of parameter interrelatedness, are displayed in Figure 4. For $N$ = 100, with no multiplicity control (NC) the power ranged from 11% to 98% as the correlation between latent factors (and consequently between parameters) increased. The average power using the original Bonferroni correction was considerably lower, ranging from 1% to 77%. As expected, the average power rates for AB1 and AB2 fell between NC and Bonferroni. At all levels of correlation between latent factors, Adjustment 1 and Adjustment 2 had higher power than Bonferroni, with the largest differences seen in the mid-range. In particular, for $r_{\text{latents}} = 0.35$, the average per-parameter power for AB1 exceeded Bonferroni by 11%, and AB2 was 28% higher than Bonferroni.

For $N$ = 400, per-parameter power quickly approached 100% (ceiling effect) as the correlations increased for all procedures, thus improvements in power were more

restricted. AB1 offered 2% to 5% higher power than Bonferroni, and the power for AB2 was 5% to 15% higher, with the largest improvement at $r_{latents} = 0.2$.

## Discussion

This investigation reconfirmed that without some form of adjustment to control the familywise Type I error rate, the probability of erroneously declaring one or more null parameters significant far exceeds the specified α-level. Furthermore, the popular Bonferroni adjustment is unacceptably conservative, resulting in an excessive penalty in statistical power. By incorporating both the number of tests in a family and the degree of correlation between parameters, two alternative adjustments have been shown to produce per-parameter power rates much closer to those obtained without multiplicity control *without inflating the familywise error rate*. More specifically, AB2 is recommended as it produced familywise error rates closest to α, and produced the greatest power of any of the familywise error controlling procedures.

With respect to the familywise error rate, one evident finding was that although the adjusted Bonferroni procedures significantly improved power, the familywise error rate was still consistently less than the nominal α. Thus, future research will hopefully explore even more powerful alternative procedures that bring the familywise error rate closer to α and the per-parameter power closer to no multiplicity control. Previous research that has attempted to consider the interrelatedness of the parameters via higher-order Bonferroni inequalities have found that the computations become excessively demanding even with very simple designs (e.g., Glaz, 1993). However, future research in this area may provide practical solutions that improve on the adjusted procedures discussed in this paper (e.g., Hoover, 1990). A related note is that in follow-up simulations to the results reported in

this paper, we found that the adjusted Bonferroni procedures (AB1 and AB2) also produced greater power than more liberal alternative Bonferroni procedures such as Holm's (1979) sequentially rejective step-down procedure, which considers the number of previously rejected hypotheses in adjusting the $\alpha_{\text{per test}}$, and thus, as discussed above, more powerful familywise error controlling alternatives to the adjusted Bonferroni methods presented in this paper will likely need to incorporate information about parameter relatedness.

It is worth noting that the described procedure need not be restricted to any particular portion of a structural model. For the purposes of this investigation Type I error control was applied specifically to the structural model, however there is no reason that multiplicity control could not also be imposed in other parts of the model. A researcher may choose to select a family of tests that includes any or all parameters (factor loadings, covariances, etc.) that are considered of particular interest within the context of the study.

To summarize, by incorporating a measure of the dependencies between SEM parameters into an adjusted Bonferroni procedure it is possible to achieve a form of multiplicity control that is far superior to the traditional Bonferroni procedure in terms of power, while still constraining the familywise Type I error rate below $\alpha$. The availability of these adjusted Bonferroni procedures allows researchers to gain substantial power without compromising control of the familywise error rate.

References


Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238-46.

Cribbie, R. A. (2007). Multiplicity control in structural equation modeling. *Structural Equation Modeling, 14*(1), 98-112.

Cribbie, R. A. (2000). Evaluating the importance of individual parameters in structural equation modeling: The need for type I error control. *Personality and Individual Differences, 29*(3), 567- 577.

Fox, J. (2006). Structural equation modeling with the sem package in R. *Structural Equation Modeling, 13*(3), 465-486.

Glaz, J. (1993). Approximate simultaneous confidence intervals. In F. M. Hoppe (ed.) *Multiple comparisons, selection, and applications in biometry* (pp. 149-166). New York: Marcel Dekker.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Hoover D. R. (1990). Subset complement addition upper bounds: An improved inclusion-exclusion method. *Journal of Statistical Planning and Inference, 24,* 195-202.

Keselman, H., Cribbie, R., & Holland, B. (2002). Controlling the rate of Type I error over a large set of statistical tests. *British Journal of Mathematical and Statistical Psychology, 55*, 27-39.

Mulaik, S. (2004, January 27). Bonferroni tests. Message posted to semnet@bama.ua.edu.

Owen, S. (2004, January 28). Bonferroni tests. Message posted to semnet@bama.ua.edu.

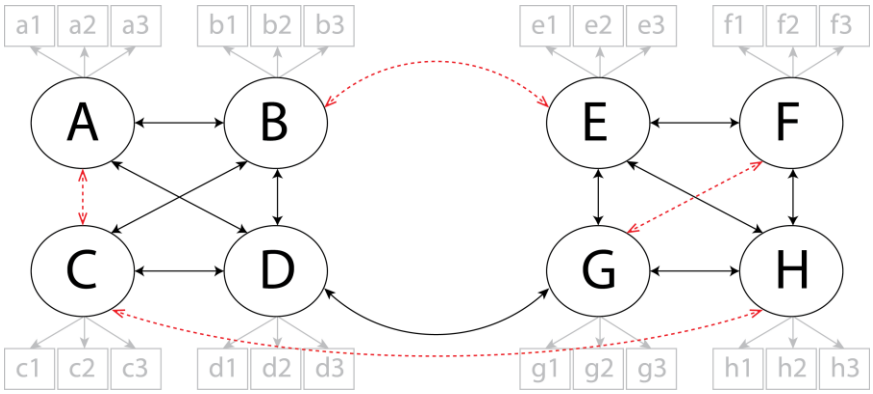
R Development Core Team (2010). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Ronis, D. L. (2002, July 25). Type I error in path analysis. Message posted to semnet@bama.ua.edu.

Steiger, J. H., & Lind, J. C. (1980, May). Statistically based tests for the number of common factors. Paper presented at the annual Spring Meeting of the Psychometric Society, Iowa City, IA.

Figure 1

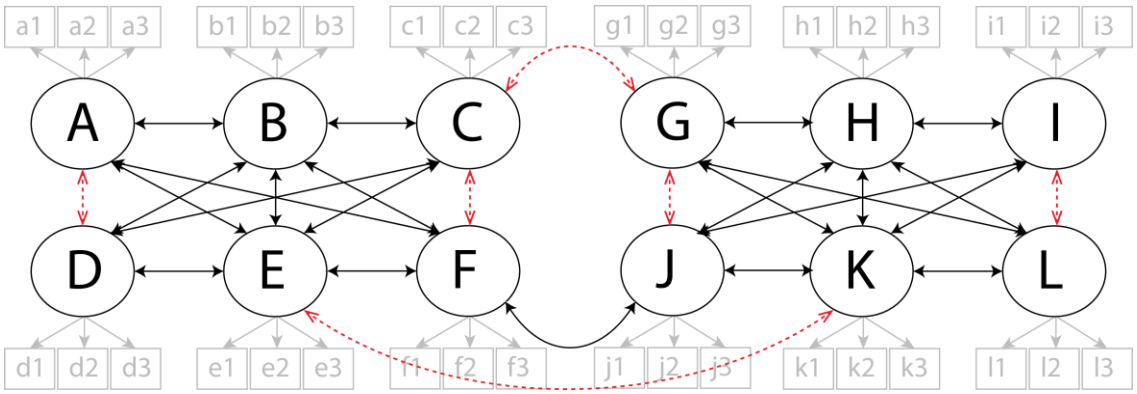
The first structural model investigated in the simulation study (Model A).




Note: Dashed lines represent null parameters (i.e., Type I errors if declared statistically significant) and solid lines represent non-null parameters (i.e., a Type II error if not declared statistically significant).

Figure 2

The second structural model investigated in the simulation study (Model B).




Note: Dashed lines represent null parameters (i.e., Type I errors if declared statistically significant) and solid lines represent non-null parameters (i.e., a Type II error if not declared statistically significant).

Figure 3

Familywise error rates for Model A (top row) and Model B (bottom row) at N = 100 and 400 at 4 different degrees of parameter interrelatedness.

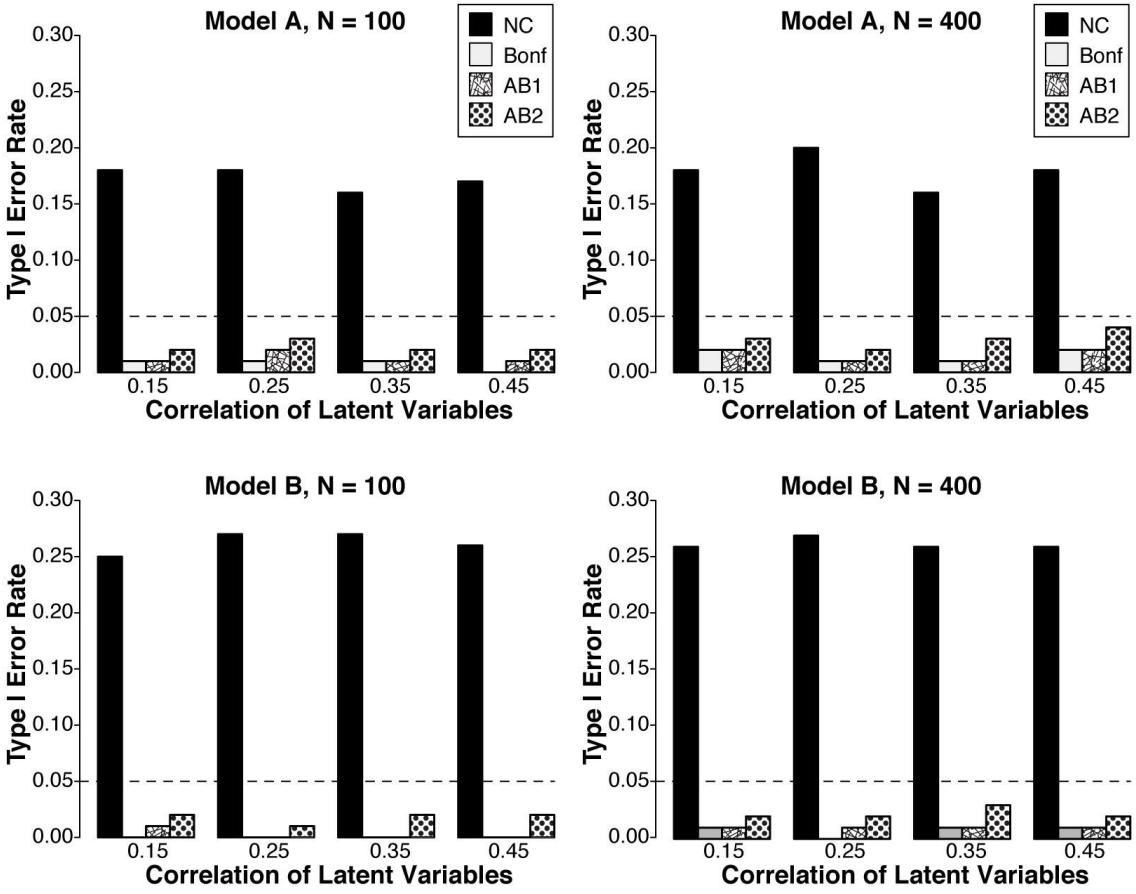


Figure 4

Per-parameter power for Model A (top row) and Model B (bottom row) at N = 100 and 400 at 4 different degrees of parameter interrelatedness.