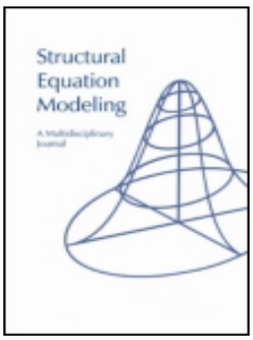

Routledge
Taylor & Francis Group

Structural Equation Modeling: A Multidisciplinary Journal

ISSN: 1070-5511 (Print) 1532-8007 (Online) Journal homepage: http://www.tandfonline.com/loi/hsem20

Multiplicity Control in Structural Equation Modeling

Robert A. Cribbie

To cite this article: Robert A. Cribbie (2007) Multiplicity Control in Structural Equation Modeling, Structural Equation Modeling: A Multidisciplinary Journal, 14:1, 98-112

To link to this article: http://dx.doi.org/10.1080/10705510709336738

STRUCTURAL EQUATION MODELING, *14*(1), 98–112

Multiplicity Control in Structural Equation Modeling

Robert A. Cribbie
York University

Researchers conducting structural equation modeling analyses rarely, if ever, control for the inflated probability of Type I errors when evaluating the statistical significance of multiple parameters in a model. In this study, the Type I error control, power and true model rates of famsilywise and false discovery rate controlling procedures were compared with rates when no multiplicity control was imposed. The results indicate that Type I error rates become severely inflated with no multiplicity control, but also that familywise error controlling procedures were extremely conservative and had very little power for detecting true relations. False discovery rate controlling procedures provided a compromise between no multiplicity control and strict familywise error control and with large sample sizes provided a high probability of making correct inferences regarding all the parameters in the model.

Researchers in the educational and behavioral sciences are increasingly turning to structural equation modeling (SEM) to answer complex multivariate hypotheses. This increase has been triggered by advances in SEM software and the availability of Web- and text-based resources for conducting SEM analyses. Although many issues surrounding the application of SEM have received dedicated attention in the literature (e.g., approximate fit indexes, estimation methods, assumption violation), research into the issue of appropriate multiplicity control when testing multiple parameters in SEM has been scarce.

When researchers are faced with evaluating the adequacy of a particular model, they are often interested in both the overall fit of the model and which of the proposed relations (parameters) in the model are (or are not) important. With respect to determining the importance of individual parameters, two alternatives have dominated the empirical literature, falling under the general categories of exploratory or confirmatory parameter investigations.

Correspondence should be addressed to Robert A. Cribbie, Department of Psychology, York University, Toronto, Ontario M3J 1P3. E-mail: cribbie@yorku.ca

Exploratory parameter investigations utilize residual statistics and sequential model modification indexes to identify which parameters to remove from or add to the model to improve the fit of the model to the sample data. This option is acceptable when the analysis is designated as exploratory and the goal of the analysis is only to derive or improve a theoretical model for future evaluation. However, researchers should be aware that (a) changes to the model are data driven and do not provide any substantive evidence about the validity of the model, (b) the probability of making decision errors is related to the number of model modifications and the criteria adopted for declaring model modifications statistically significant (Green & Babyak, 1997; Hancock, 1999; MacCallum, 1986; MacCallum, Roznowski, & Necowitz, 1992), and (c) specification searches can be unreliable detectors of specification errors (e.g., Hutchinson, 1993; Kaplan, 1988; MacCallum, 1986; MacCallum et al., 1992; Silvia & MacCallum, 1988).

Confirmatory parameter investigations (assuming a satisfactory fit of the model to the sample data) evaluate the significance of each hypothesized parameter at a specified significance level (α), declaring statistically significant parameters important and nonsignificant parameters unimportant, within the framework of the model. This strategy is acceptable only if the researcher is willing to assume that no Type I error inflation will arise when multiple hypotheses are tested.

The current state of multiplicity control in confirmatory SEM can be summarized very easily: There is no multiplicity control! However, that is not to say that there is no discussion of the issue. A review of the archives of SEMNET, the online SEM network group, produced several comments regarding multiplicity control in SEM. For example:

- "I am wondering if you should use a Bonferroni correction procedure for interpreting critical ratios (also called *t* tests). For example, if 10 estimates are produced, perhaps you should use a significance level of .005 for any one parameter" (Burns, 1996).
- "Many students regrettably pick out the 'significant' results and report only those. One pragmatic approach is to apply the Bonferroni correction when reporting only N out of M tests applied" (Reese, 2001).
- "My statistics professor on my dissertation committee asked me how does AMOS control for Type I error (likelihood of finding significant relationships when running multiple analyses on the same data). He asked is it doing a Bonferroni adjustment, or what is it doing? I will need to address this in my dissertation revision" (Moynihan, 2002).
- "I was talking to my supervisors (who have never used SEM before) and they argued that you could expect 1 in 20 correlations between variables to be significant and if you have more than 20 paths that you could expect that one of those paths would be significant even if there was no correlation between the variables. They wondered if it makes sense to apply a Bonferroni correction

> on the *p* values of your estimated parameters to account for this possibility" (Van der Heijden, 2005).

Popular responses against multiplicity control in SEM research usually fall under one of two categories: (a) the issues surrounding multiplicity control in SEM are identical to the issues raised in other forms of correlational analysis (where only very rarely is multiplicity control invoked when multiple parameters or correlations are evaluated); or (b) the parameters in SEM models are correlated so multiplicity adjustments would be too conservative (e.g., Mulaik, 2004; Owen, 2004; Ronis, 2002). However, the arguments in support of multiplicity control are extremely convincing. First, there have been many (mostly ignored) warnings concerning multiplicity control in correlational research (e.g., Collis & Rosenblood, 1985; Crosbie, 1986; Cudeck & O'Dell, 1994; Larzelere & Mulaik, 1977) and yes, the same issues apply to SEM research. Second, the hypotheses tested in SEM are often more confirmatory in nature, and regarded as much more definitive, than other correlational analyses and therefore warrant significantly more attention. Third, the fact that parameters in a model are often highly intercorrelated does affect the conservativeness of Type I error controlling procedures, but because researchers are typically unaware of the degree to which parameters within their model are correlated, and the correlations between parameter estimates decrease when a model is well identified, researchers must be conscious of the possible risk of falsely declaring parameters significant. MacCallum (1995) added that although models with several correlated parameters tend to fit the data well, they are often not disconfirmable and thus make finding a good fit meaningless. Finally, in a preliminary study of the impact of interpreting the significance of several parameters in a structural model on the inflation of Type I errors, Cribbie (2000) concluded that the Type I error inflation was too extreme to ignore and that some form of multiplicity control was necessary.

An important issue that frequently arises in the application of multiplicity control is what family or set of parameters in which to apply the control. The goal of specifying a family of hypotheses is to select a set that is not so large that it is impossible to ever reject any hypothesis, yet is not so small that it does not provide adequate control of Type I errors. Further, a family should consist of those tests that are related in terms of their intended use (Hochberg & Tamhane, 1987). In this research, Type I error control was applied over all parameters in the structural model, which often represent the important hypotheses within the model. However, there is no reason that multiplicity control could not also be imposed in other parts of the model, and may be extremely relevant in research where the measurement model is of utmost importance (e.g., confirmatory factor analysis).

With respect to multiplicity control, there are several different units of analysis (i.e., error rates) that have been proposed and that vary in how strictly they

control the rate of Type I errors. Although the majority of discussion in the literature has focused on the familywise error rate versus no multiplicity control (e.g., Ryan, 1959; Toothaker, 1991; Tukey, 1953), other error rates, such as the false discovery rate (FDR; Benjamini & Hochberg, 1995), have also been proposed. When no multiplicity control is imposed, all tests of significance are conducted at α (i.e., the per-test Type I error rate). The advantages of imposing no control include consistency (regardless of the number of tests being conducted, the Type I error rate for each test is held constant) and power, and the disadvantage is that there is a potential increase in the overall Type I error rate when multiple tests of significance are performed. The familywise error rate is defined as the probability of falsely rejecting one or more hypotheses in a family of hypotheses. The main advantage of familywise error control is that the probability of committing a Type I error does not increase as more tests of significance are conducted, whereas the disadvantage is that per-test power can become severely deflated as the number of hypothesis tests increases. Finally, Benjamini and Hochberg (1995) presented a compromise between no multiplicity control and strict familywise error control, namely the FDR. The FDR is defined as the expected ratio (Q) of the number of erroneous rejections (V) to the total number of rejections (R = V + S), where S represents the number of true rejections (Benjamini, Hochberg, & Kling, 1994). Therefore, E (Q) = E (V / (V + S)) = E (V / R). The relation between FDR control and other error rates was summarized by Benjamini et al. (1994). If all null hypotheses are true, the FDR equals the familywise error rate. On the other hand, if some of the null hypotheses are false, the FDR is less than the familywise error rate, which can result in a significant increase in power for a procedure that controls the FDR. Keselman, Cribbie, and Holland (2002) demonstrated that the FDR procedure provides excellent power (relative to familywise error controlling procedures) in experiments with large family sizes (e.g., testing all correlations in a 16 × 16 matrix), while still providing acceptable Type I error control.

Therefore, the purpose of this investigation is to explore the application of multiplicity control within the framework of SEM and compare available methods for multiplicity control with respect to Type I error control and power.

METHOD

A Monte Carlo study was used to compare the Type I error control, per parameter power, and true model rates of the no multiplicity control approach with that of two familywise error controlling procedures and two FDR controlling procedures.

Multiple Testing Procedures

Bonferroni (Bonf). In this well-known procedure, the p values corresponding to the k parameter estimates are compared to an $\alpha_{pp} = \alpha/k$, where α_{pp} is the per-parameter Type I error rate and α is a predetermined, acceptable, familywise error rate. The Bonferroni method controls the familywise error rate at α by declaring any parameter with $p \leq \alpha_{pp}$ significant.

Hochberg's (1988) sequentially acceptive step-up Bonferroni (Hoch). In this familywise error controlling procedure, the p values corresponding to the k parameter estimates are ordered from smallest to largest. Then, for any $I = k, k - 1, \ldots, 1$, if $p_i \leq \alpha/(k - I + 1)$, the Hochberg procedure rejects all parameters associated with $p_{i'}$ $(I' \leq I)$. Therefore, one begins by testing the largest p value, p_k, and declares all parameters significant if $p_k \leq \alpha$. If $p_k > \alpha$ then the parameter associated with p_k is declared nonsignificant and one proceeds to compare p_{k-1} to $\alpha/2$. If $p_{k-1} \leq \alpha/2$ then all parameters associated with p_i $(I = k - 1, \ldots, 1)$ are declared significant, but if $p_{k-1} > \alpha/2$ then the parameter associated with p_{k-1} is declared nonsignificant and one proceeds to compare p_{k-2} to $\alpha/3$, and so on.

Benjamini and Hochberg's (1995) false discovery rate controlling step-up Bonferroni (FDR). Like Hochberg's step-up procedure, the p values corresponding to the k parameter estimates are ordered from smallest to largest. Then, for any $I = k, k - 1, \ldots, 1$, if $p_i \leq \alpha(I/k)$, the FDR procedure rejects all parameters associated with p_i $(I' \leq I)$. Therefore, one begins by testing the largest p value, p_k, and declares all parameters significant if $p_k \leq \alpha$. If $p_k > \alpha$ then the parameter associated with p_k is declared nonsignificant and one proceeds to compare p_{k-1} to $\alpha(k - 1/k)$. If $p_{k-1} \leq \alpha(k - 1/k)$ then all parameters associated with p_i $(I = k - 1, \ldots, 1)$ are declared significant, but if $p_{k-1} > \alpha(k - 1/k)$ then the parameter associated with p_{k-1} is declared nonsignificant and one proceeds to compare p_{k-2} to $\alpha(k - 2/k)$, and so on. Benjamini and Yekutieli (2001) provided evidence that this procedure provides strong FDR control under the type of parameter dependencies encountered in latent variable models.

Benjamini and Yekutieli's (2001) FDR (FDR–BY). Benjamini and Yekutieli (2001) proposed a modification to the original FDR procedure that would be more conservative and control the FDR across any parameter dependence structure (referred to here as the FDR–BY). Like the original FDR procedure, the p values corresponding to the k parameter estimates are ordered from smallest to largest. Then, for any $I = k, k - 1, \ldots, 1$, if $p_i \leq \alpha[I / (k \Sigma_r 1/r), r = 1, \ldots, k$, the FDR–BY procedure rejects all parameters associated with $p_{i'}$ $(I' \leq I)$.

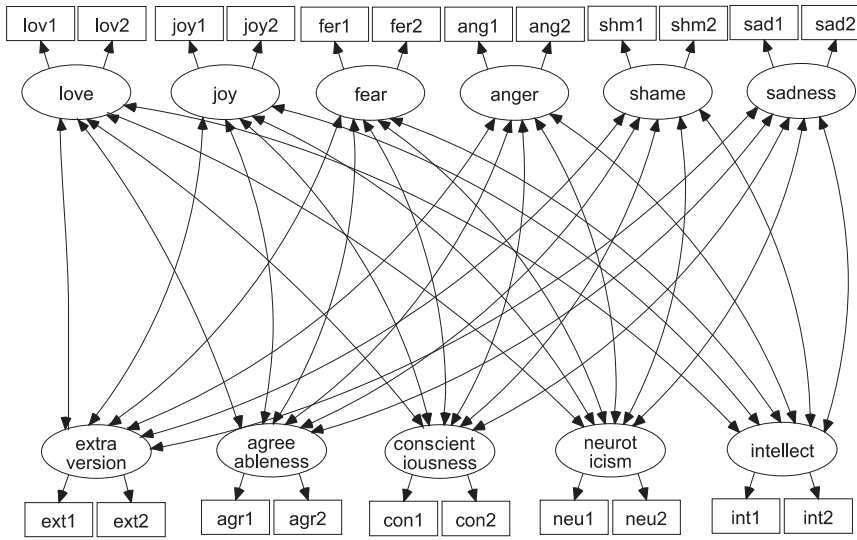

FIGURE 1 The first theoretical model used in the Monte Carlo study (Model A), based on a study by Trierweiler, Eid, and Lischetzke (2002).

The models investigated in this article were derived from studies by Trierweiler, Eid, and Lischetzke (2002, Model A; see Figure 1) and Dunkley, Zuroff, and Blankstein (2003, Model B; see Figure 2). In the Trierweiler et al. (2002) study, the authors were interested in exploring relations between emotional expressions (love, joy, fear, anger, shame, sadness) and the Big Five personality dimensions (extraversion, agreeableness, conscientiousness, neuroticism, intellect). In the Dunkley et al. (2003) study, the authors were interested in exploring predictors of negative and positive affect, including hassles, avoidant coping, perceived social support, self-critical perfectionism, personal standards perfectionism, event stress, and problem-focused coping. Empirical models were utilized to provide a theoretical framework for exploring issues related to assessing multiple hypotheses, and these particular models were selected because there were a moderate to large number of constructs that would likely be familiar to most readers. Model A had 179 *df* with 30 hypothesis tests in the structural model and Model B had 150 *df* with 20 hypothesis tests in the structural model. 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 true model rate was evaluated as the proportion of simulations in which the true underlying model configuration was recovered (i.e., all nonzero parameters were declared statistically significant and all null parameters were declared not statisti-

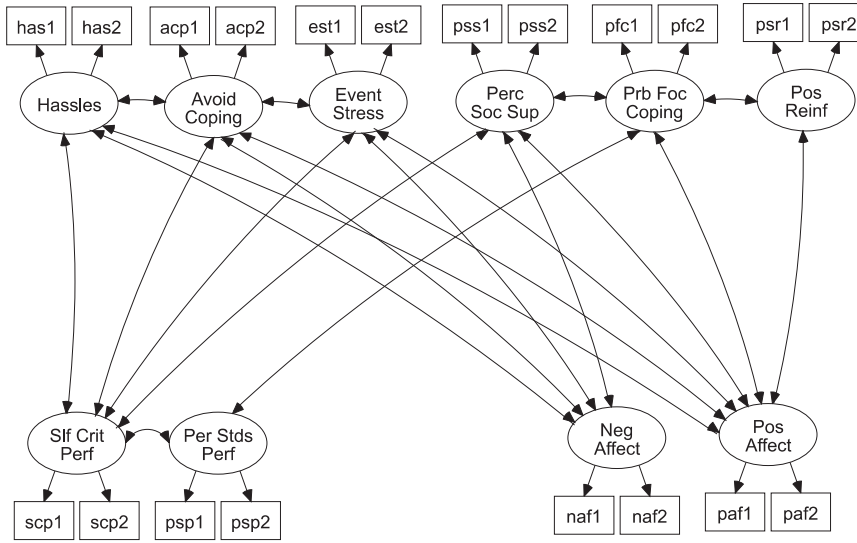

FIGURE 2 The second theoretical model used in the Monte Carlo study (Model B), based on a study by Dunkley, Zuroff, and Blankstein (2003). Note: avoid coping = avoidant coping; per soc sup = perceived social support; slf crt perf = self-critical perfectionism; per stds perf = personal standards perfectionism; prb foc coping = problem-focused coping.

cally significant). Cribbie (2003) and Cribbie and Keselman (2003) previously used the true model rate to compare the Type I error control and power of multiple comparison procedures in a mean comparison framework. Although the true model rate is an extremely conservative criterion, it is recommended as a measure of the performance of multiple testing procedures because it simultaneously investigates Type I error control and power.

Three other variables were investigated in this study: (a) sample size ($N = 200$ and $N = 1{,}000$), (b) type of misspecification, and (c) number of misspecifications in the model (6 or 12 for Model A; 5 or 10 for Model B). The sample sizes were selected to be representative of those encountered by applied researchers. Misspecifications were established by estimating (freeing) paths between variables in the model that are not related in the true model. In addition to a no misspecification condition, two forms of misspecification were investigated in this study: (a) dependent misspecification, where all misspecified paths originate from one (in the case of 5 or 6 misspecifications) or two (in the case of 10 or 12 misspecifications) of the latent variables; or (b) independent misspecification, where the misspecified paths were as unrelated as possible. In deriving an implied covariance matrix, population factor variances were set at 1.0, interfactor covariances were set at .18, factor loadings were set to .80, and error variances were set to .36 (resulting in unit variance for the observed variables). Misspecifications (following the format de-

scribed earlier) were established by fixing the covariances between observed variables loading on separate, but theoretically correlated, factors to 0 in the model specified covariance matrix. To summarize the design of the study, there are two separate models, two sample size conditions, two types of misspecification (in addition to a case where no misspecifications were present), and two conditions for the number of misspecifications.

SEM analyses were performed using SAS PROC CALIS (SAS Institute, 1999) and preliminary analyses were conducted to determine the fit of the models to the data using the comparative fit index (CFI) and root mean square error of approximation (RMSEA). One thousand replications were performed for each condition using a nominal significance level of .05.

RESULTS

Fit Indexes

The fit indexes under each of the experimental conditions are presented in Table 1. The fit of the model to the sample data was excellent for both models under all conditions investigated in this study, with mean CFI values greater than .99 and RMSEA values less than .02. An interesting finding was that the fit of the models remained excellent even when several of the parameters in the structural model were misspecified.

TABLE 1
Average Fit Indexes for Each Model as a Function of Sample Size, Misspecification Structure, and Number of Misspecifications

	Sample Size			
	N = 200		*N = 1,000*	
Type (and Number) of Misspecifications	*CFI*	*RMSEA*	*CFI*	*RMSEA*
Model A				
No misspecification	.991	.013	.999	.005
Dependent (6)	.992	.012	.999	.004
Independent (6)	.993	.012	.999	.005
Dependent (12)	.994	.010	.999	.003
Independent (12)	.993	.011	.999	.004
Model B				
No misspecification	.993	.011	.999	.005
Dependent (5)	.993	.011	.999	.005
Independent (5)	.993	.011	.999	.004
Dependent (10)	.994	.010	.999	.004
Independent (10)	.993	.010	.999	.004

Note. CFI = comparative fit index; RMSEA = root mean square error of approximation.

Familywise Error Rates

Familywise error rates for each of the multiple testing procedures, under each of the testing conditions, are presented in Table 2. As expected, the familywise error rates when no multiplicity control was imposed significantly exceeded the per-parameter alpha of .05, and were larger than the rates for any of the remaining procedures. The familywise rates when no multiplicity control was imposed ranged from .205 for 6 dependent misspecifications and $N = 1{,}000$ to .616 for 12 independent misspecifications and $N = 200$. The familywise error rates for the FDR procedure were greater than for the remaining procedures, with the Bonf approach having the most conservative rates. The FDR–BY procedure was extremely conservative with a small sample size, but had larger rates than

TABLE 2
Familywise Error Rates for Each Multiple Testing Procedure as a Function of Sample Size, Misspecification Structure, and Number of Misspecifications

Type (and Number) of Misspecifications	*NC*	*Bonf*	*Hoch*	*FDR*	*FDR–BY*
$N = 200$					
Model A					
Dependent (6)	.243	.007	.007	.064	.005
Independent (6)	.348	.011	.012	.104	.007
Dependent (12)	.287	.007	.009	.069	.004
Independent (12)	.616	.069	.073	.213	.043
Model B					
Dependent (5)	.309	.016	.017	.077	.011
Independent (5)	.331	.016	.021	.084	.011
Dependent (10)	.570	.039	.039	.109	.012
Independent (10)	.582	.045	.049	.110	.020
$N = 1{,}000$					
Model A					
Dependent (6)	.205	.006	.029	.178	.041
Independent (6)	.295	.005	.042	.249	.060
Dependent (12)	.306	.011	.030	.197	.052
Independent (12)	.468	.012	.032	.335	.072
Model B					
Dependent (5)	.228	.009	.038	.171	.042
Independent (5)	.253	.009	.046	.202	.055
Dependent (10)	.454	.019	.034	.268	.068
Independent (10)	.436	.023	.046	.259	.076

Note. NC = no familywise error control; Bonf = Bonferroni; Hoch = Hochberg; FDR = false discovery rate; FDR–BY = Benjamin & Yekutieli's conservative FDR; Dependent = misspecifications from the same latent variable(s); Independent = misspecifications not all from the same latent variable(s).

the familywise error controlling procedures (Bonf, Hoch) with a large sample size.

Per-Parameter Power

Per-parameter power rates for each of the multiple testing procedures, under each of the testing conditions, are presented in Table 3. Per-parameter power rates when no multiplicity control was imposed were, as expected, larger than for any of the remaining procedures in any condition, ranging from .242 to .527 for $N = 200$ and

TABLE 3
Per-Parameter Power Rates for Each Multiple Testing Procedure as a Function of Sample Size, Misspecification Structure, and Number of Misspecifications

Type (and Number) of Misspecifications	*NC*	*Bonf*	*Hoch*	*FDR*	*FDR–BY*
$N = 200$					
Model A					
No misspecification	.527	.083	.089	.353	.074
Dependent (6)	.359	.054	.056	.195	.039
Independent (6)	.471	.071	.075	.240	.049
Dependent (12)	.242	.035	.036	.100	.021
Independent (12)	.383	.055	.057	.153	.032
Model B					
No misspecification	.455	.080	.086	.251	.059
Dependent (5)	.427	.072	.076	.176	.045
Independent (5)	.429	.073	.077	.180	.045
Dependent (10)	.398	.067	.069	.012	.034
Independent (10)	.406	.070	.073	.130	.038
$N = 1{,}000$					
Model A					
No misspecification	.996	.911	.995	.996	.980
Dependent (6)	.884	.640	.680	.714	.686
Independent (6)	.995	.886	.956	.993	.969
Dependent (12)	.834	.486	.509	.566	.519
Independent (12)	.991	.854	.906	.984	.945
Model B					
No misspecification	.992	.887	.990	.992	.966
Dependent (5)	.987	.861	.930	.982	.944
Independent (5)	.988	.854	.926	.983	.942
Dependent (10)	.982	.829	.870	.966	.901
Independent (10)	.984	.824	.866	.968	.895

Note. NC = no familywise error control; Bonf = Bonferroni; Hoch = Hochberg; FDR = false discovery rate; FDR–BY = Benjamin & Yekutieli's conservative FDR; Dependent = misspecifications from the same latent variable(s); Independent = misspecifications not all from the same latent variable(s).

.834 to .996 for $N = 1{,}000$. Per-parameter power rates for the FDR procedure, ranging from .100 to .353 for $N = 200$ and .566 to .996 for $N = 1{,}000$, were larger than for the remaining familywise error controlling procedures (Bonf, Hoch), although the differences were less pronounced in the large sample size condition where ceiling effects limited the amount of variability in the rates. The Hoch procedure was slightly more powerful than the Bonf procedure in all conditions investigated. Per-parameter power rates for the FDR–BY procedure were less than those for the FDR, but larger than for the Bonf and Hoch familywise error controlling procedures with a large sample size. For Model A, power rates when there were dependent misspecifications were significantly depressed relative to the rates for independent misspecifications, whereas for Model B (where the parameter structure is less defined) power rates when there were dependent misspecifications were very similar to rates when there were independent misspecifications.

True Model Rates

True model rates for each of the multiple testing procedures for $N = 1{,}000$ are presented in Table 4. The true model rates investigate the performance of multiple testing procedures by simultaneously incorporating both Type I error control and power (in other words to detect the true model means that no Type I or Type II errors were committed). True model rates for $N = 200$ were 0, regardless of

TABLE 4
True Model Rates for Each Multiple Testing Procedure for $N = 1{,}000$ by Misspecification Structure and Number of Misspecifications

Type (and Number) of Misspecifications	*NC*	*Bonf*	*Hoch*	*FDR*	*FDR–BY*
Model A					
No misspecification	.874	.046	.874	.874	.530
Dependent (6)	.443	.057	.278	.451	.335
Independent (6)	.624	.054	.342	.636	.446
Dependent (12)	.204	.056	.149	.273	.243
Independent (12)	.450	.062	.191	.498	.353
Model B					
No misspecification	.850	.089	.850	.850	.515
Dependent (5)	.631	.114	.366	.637	.418
Independent (5)	.625	.108	.349	.628	.401
Dependent (10)	.461	.146	.258	.518	.343
Independent (10)	.507	.144	.252	.549	.331

Note. NC = no familywise error control; Bonf = Bonferroni; Hoch = Hochberg; FDR = false discovery rate; FDR–BY = Benjamin & Tekutieli's conservative FDR; Dependent = misspecifications from the same latent variable(s); Independent = misspecifications not all from the same latent variable(s).

the number or type of misspecifications or the multiple testing procedure utilized and are therefore not presented. For $N = 1{,}000$ and no misspecifications, the true model rates of the no multiplicity control, FDR, and Hoch procedures were equal (.874). When misspecifications were present, the true model rates of the FDR procedure were larger than the rates of any of the other procedures across all conditions, with rates ranging from .273 to .636 for Model A and .518 to .637 for Model B. The true model rates in the no multiplicity control condition were slightly less than that of the FDR procedure, and the true model rates of the FDR–BY procedure were larger than that of the Hoch and Bonf procedures. True model rates for the Bonf procedure were extremely low, reaching a maximum of .146.

DISCUSSION

Researchers conducting SEM analyses rarely (or possibly never) impose any type of multiplicity control when evaluating the significance of multiple parameters. However, past research has shown that even when dependencies among the parameters exist, familywise error rates are expected to be inflated when multiple null parameters are estimated (e.g., Larzelere & Mulaik, 1977). This article looked at the effects of evaluating the statistical significance of multiple parameters in the structural model and the results confirm that Type I error inflation becomes extreme when no multiplicity control is imposed. The Type I error inflation is slightly reduced when dependencies among parameters exist, although the rates still greatly exceed the nominal alpha level. Is the rate of Type I error inflation excessive? That is a difficult question, but I believe the point is that if a researcher believes that his or her probability of erroneously declaring any parameter significant is alpha (say .05), then any inflation of the Type I error rate above .05 would be problematic and for most researchers the .20 to .60 probabilities of erroneously declaring a parameter significant reported in this article would likely be alarming.

This research confirms that the original FDR procedure (Benjamini & Hochberg, 1995) provides a compromise position between no multiplicity control and strict familywise error control, providing more power than familywise error controlling procedures but more Type I error control than when no multiplicity control is imposed. Further, the results of this investigation demonstrate the extreme conservativeness of the Bonferroni procedure relative to the remaining procedures. In fact, besides ease of computation, there is no reason to recommend the Bonferroni procedure even if strict Type I error control is desired, given that other familywise error controlling procedures (e.g., the Hochberg procedure used in this study) can provide good Type I error control and can be significantly more powerful than the Bonferroni procedure. The FDR procedure due to Benjamini and

Yekutieli (2001) was extremely conservative with small sample sizes, but provided good Type I error control with larger sample sizes and was more powerful than the Bonferroni or Hochberg procedures. In fact, although the procedure is not designed to control the familywise error rate at alpha, the familywise rates never exceeded .076 in any of the conditions investigated in this study.

An important component to this investigation was an evaluation of the true model rates of the multiple testing procedures. The true model rates, or the probability of making correct statistical decisions regarding all parameters in the structural model, were largest for the more liberal Type I error controlling strategies. More specifically, the true model rates were highest when the original Benjamini and Hochberg (1995) FDR procedure was utilized (although in most cases the rates were only slightly greater than that of the no familywise error control approach). The true model rates of the more conservative FDR controlling procedure due to Benjamini and Yekutieli (2001) were very respectable when sample sizes were large, especially considering the strict Type I error control that was observed. The true model rates also confirm that controlling the familywise error rate with the conservative Bonferroni correction is not recommended and make it highly unlikely that a researcher will uncover all of the true relations in the model.

An important limitation of this study was that only two models were considered. It will be important in future research to explore how the results of this study extend to other types of models, which can obviously vary considerably in both the nature of the model as well as the number of parameters estimated. A reviewer of this article also pointed out that the reliability of the indicators was not varied, and could possibly affect the results. Future studies should also investigate the effects of indicator reliability on parameter error rates.

To summarize, the results of this investigation highlight that the Type I error inflation that occurs when multiple parameters are investigated in SEM can become extreme, but also point out that the best methods for maximizing the probability of making correct inferences regarding all parameters in the model are the most liberal methods that result in the highest rates of Type I error. So how does this help in making recommendations to researchers regarding multiplicity control in SEM? First, consider the response of Mulaik (2004) to a researcher enquiring about whether to adopt Bonferroni control when evaluating the significance of multiple parameters in a model:

> You do not ordinarily take an alpha level and divide it by the number of dependent tests to perform to get an alpha-per-test. That could be too conservative. What the alpha-per-test should be based on is an initial risk expressed as a probability with which one would be willing to accept making at least one type I error among the series of tests. The alpha-per-test is then the result of dividing alpha familywise by the number k of tests to perform in the family or series. There is no absolute rule by which one would select the alpha familywise. But it should be larger than the usual

.05. Given many tests with an alpha-per-test of .05/k, it would be unduly conservative in favor of accepting the null hypothesis over the series.

Mulaik's response that a Bonferroni correction would be too conservative mirrors the results of this study. Further, Mulaik's suggestion that a familywise error rate should be selected based on a risk with which one would be willing to accept making at least one Type I error over all parameters is in line with the goals of this article, but what familywise rate would we recommend to researchers (.10, .20, etc.)?

Given the findings of this research it is recommended that, instead of focusing on selecting an appropriate elevated familywise error rate, that researchers maintain alpha at the desired level (e.g., .05), but adopt control of the FDR. If maximizing power is important then the original FDR procedure due to Benjamini and Hochberg (1995) would be recommended, but if more strict Type I error control is necessary the more conservative Benjamini and Yekutieli (2001) FDR controlling procedure would be recommended. Either of these methods will provide researchers with more Type I error control than when no multiplicity control is imposed, but more power than when a familywise error controlling procedure is adopted.

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 B, 57,* 289–300.

Benjamini, Y., Hochberg, Y., & Kling, Y. (1994). *False discovery rate controlling procedures for pairwise comparisons.* Unpublished manuscript.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics, 29,* 1165–1188.

Burns, D. (1996, April 17). A quick question. Message posted to semnet@bama.ua.edu

Collis, B. A., & Rosenblood, L. K. (1985). The problem of inflated significance when testing individual correlations from a correlation matrix. *Journal for Research in Mathematics Education, 16,* 52–55.

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,* 567–577.

Cribbie, R. A. (2003). Pairwise multiple comparisons: New yardstick, new results. *The Journal of Experimental Education, 71,* 251–265.

Cribbie, R. A., & Keselman, H. J. (2003). Pairwise multiple comparisons: A model comparison approach versus stepwise procedures. *British Journal of Mathematical and Statistical Psychology, 56,* 167–182.

Crosbie, J. (1986). A Pascal program to perform the Bonferroni multistage multiple-correlation procedure. *Behavior Research Methods, Instruments and Computers, 18,* 327–329.

Cudeck, R., & O'Dell, L. L. (1994). Applications of standard error estimates in unrestricted factor analysis: Significance tests for factor loadings and correlations. *Psychological Bulletin, 115,* 475–487.

Dunkley, D. M., Zuroff, D. C., & Blankstein, K. R. (2003). Self-critical perfectionism and daily affect: Dispositional and situational influences on stress and coping. *Journal of Personality and Social Psychology, 84,* 234–252.

Green, S. B., & Babyak, M. A. (1997). Control of Type I errors with multiple tests of constraints in structural equation modeling. *Multivariate Behavioral Research, 32,* 39–51.

Hancock, G. R. (1999). A sequential Scheffé-type respecification procedure for controlling Type I error in exploratory structural equation model modification. *Structural Equation Modeling, 6,* 158–168.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75,* 800–802.

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures.* New York: Wiley.

Hutchinson, S. R. (1993). Univariate and multivariate specification search indices in covariance structure modeling. *Journal of Experimental Education, 61,* 171–181.

Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivariate Behavioral Research, 23,* 69–86.

Keselman, H. J., Cribbie, R. A., & 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.

Larzelere, R. E., & Mulaik, S. A. (1977). Single-sample tests for many correlations. *Psychological Bulletin, 84,* 557–569.

MacCallum, R. C. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin, 100,* 107–120.

MacCallum, R. C. (1995). Model specification: Procedure, strategies, and related issues. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 16–36). Newbury Park, CA: Sage.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111,* 490–504.

Moynihan, L. (2002, July 23). Type I error in path analysis. Message posted to semnet@bama.ua.edu

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

Reese, R. A. (2001, June 1). Reporting significance level. Message posted to semnet@bama.ua.edu

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

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

SAS Institute. (1999). *SAS/STAT user's guide* (version 6, 4th ed.). Cary, NC: Author.

Silvia, E. S. M., & MacCallum, R. C. (1988). Some factors affecting the success of specification searches in covariance structure modeling. *Multivariate Behavioral Research, 23,* 297–326.

Toothaker, L. E. (1991). *Multiple comparisons for researchers*. Newbury Park, CA: Sage.

Trierweiler, L. I., Eid, M., & Lischetzke, T. (2002). The structure of emotional expressivity: Each emotion counts. *Journal of Personality and Social Psychology, 82,* 1023–1040.

Tukey, J. W. (1953). *The problem of multiple comparisons.* Unpublished manuscript, Princeton University, Department of Statistics.

Van der Heijden, G. (2005, June 8). Bonferroni correction. Message posted to: semnet@bama.ua.edu.