# The problem of allowing correlated errors in structural equation modeling: concerns and considerations

**Richard HERMIDA**[*]
George Mason University, 3575 Owasso Street, Shoreview, MN, USA, 55126

## Abstract

*Results of structural equation models may be negated by inappropriate methodological procedures. Model fit is known to be improved by the addition of pathways. Some pathways are added due to modification indices. These a-theoretical pathways will improve model fit at the expense of theory and reduction in parameter value replication. Furthermore, some additions to the model like correlating measurement errors are usually theoretically unjustifiable. This quantitative review examines the frequency of correlating measurement errors and examines the reasons, if any, for having these pathways in the model. Additionally, this quantitative review examines the consequences of correlating measurement errors in structural equation modeling.*

**Keywords:** *Structural Equation Modeling, Confirmatory Factor Analysis, Measurement, Research Methods, Statistics.*

## Introduction

The use of structural equation modeling (SEM) to test theoretical models has increased dramatically over the past 25 years in fields such as psychology, sociology, economics, marketing, and even behavior genetics (Markon & Kruger, 2004). While advances in statistical software packages have made it easier than ever for researchers to employ structural equation models (MacCallum & Austin, 2000), dissemination of best practices in SEM has arguably not kept pace. For example, there seems to confusion regarding the practice of allowing measurement errors to correlate in order to improve model fit. Many authors have cautioned against this practice (see Brannick, 1995; Cliff, 1983; Cortina, 2002; Gerbing & Anderson, 1984; Kaplan, 1989; Kaplan, 1990; MacCallum, 1986; MacCallum, Roznowski, & Waller, 1992; Shah & Goldstein, 2006; Steiger, 1990; Tomarken & Waller, 2003) for a variety of methodological reasons.

Other authors have attempted to identify the situations in which it is appropriate. Landis, Edwards, and Cortina (2009) argue that estimation of measurement errors in SEM is only appropriate when correlations amongst measurement errors are unavoidable. Such situations include when multiple measures of the same construct are used in longitudinal research, or when indicator variables share components. However, when measurement errors are allowed to correlate based on post hoc specification searches, a theoretical justification for engaging in this practice is lacking and the corresponding research moves from a confirmatory analysis to an exploratory analysis. Indeed, researchers have argued that such practices may only increase model fit indices because of capitalization on chance. Freeing the paths between measurement errors is particularly problematic in post hoc specification searches given that the correlations between errors may indicate that the model is misspecified; that correlating measurement errors based on post hoc modification may actually mask the underlying structure of the data; and that there is no theoretically defensible reason for allowing measurement errors to correlate based on post hoc modifications (Anderson & Gerbing, 1984; Landis et al., 2009).

Given the lack of consistent guidelines, researchers continue to engage in this practice without justification. Consequently, it is important to review the state of this practice as there is a high potential for misguided conclusions from structural equation models which could possibly have acceptable fit statistics due to estimation of correlations among measurement errors. Additionally, formal guidelines can only be established once it is understood how extensive this problem is and the reasons for which researchers allow measurement errors to correlate. As a result, the purpose of this article is to conduct a methodological review of the current practices of correlating SEM measurement errors at both the measurement and structural levels. Specifically, this review will examine the extent to which this practice is conducted within psychology and other related disciplines. It will also determine if this practice differs by journal impact rating; will explore the degree to which researchers

[*] Corresponding author: rhermida3@gmail.com

provide citations or explanations for allowing measurement errors to correlate; will detail what explanations have been provided; and will assess whether explanations have been provided a priori or post hoc in manuscripts. Finally, this review will uncover the amount, on average, that model fit is increased by allowing measurement errors to correlate as well as determine if this practice is influenced by model complexity.

### Overview

This paper is organized as follows: First, a description of fit indices and specification searches in SEM is provided. Second, a presentation of the problems associated with allowing indicator residuals to correlate when testing both measurement and structural models is provided. Third, a brief review of past studies assessing the practice of allowing measurements errors to correlate is provided. Last, Iwill describe two factors that Ihypothesize to be related to the practice of error correlation: model complexity and journal impact. Iconclude the introduction with a summary of the unique contributions of this study. Current practices of allowing measurement errors to correlate are subsequently examined and meta-analyzed. Results of the examination are then discussed, along with the implications the current studies' findings have on the interpretation of results put forth from SEM researchers.

### SEM and Model Fit Indices

In SEM, global model fit statistics are computed based on the pathways not estimated in the hypothesized model (Jöreskog, 1969). The addition of model paths therefore, will decrease the discrepancy between the observed and reproduced variance-covariance matrices. Commonly reported model fit indices include the chi-square statistic, root mean square error of approximation (RMSEA), comparative fit index (CFI), and the normed fit index (NFI), although many other global fit indices are available in SEM data analysis programs. Additionally, each individual path in the model can be examined for magnitude and statistical significance.

When evaluating the measurement model portion of a theoretical model in SEM, two distinct parameter estimates are generated for each indicator variable (Landis et al, 2009). The first, true score common variance, is the shared variability an indicator has with other indicators of the same latent variable (Maruyama, 1998).Thus, the term true score denotes that this estimate represents the underlying latent variable that the indicator variables were hypothesized to measure. The term common variance implies that the variance derived is shared by all indicators of the same latent trait. The factor loading of a given indicator on an underlying latent variable should be large and statistically significant to the extent that the indicator has a large degree of true score common variance.

The second estimate is a residual term into which all other unique sources of variance go. Residual terms can be broken down into true score unique variance, defined as the systematic variance of an indicator that is uncorrelated with the variance of other indicators, and error variance, or the unsystematic variability of an indicator (Maruyama, 1998). When estimating the parameters of measurement models, it is typical for the covariances among measurement errors to be fixed at zero given that residual variances are defined by their uniqueness to each indicator (Landis et al, 2009).

### Specification Searches

A common procedure in SEM when a model has inadequate fit is to perform a specification search. Specification searchers are post hoc explorations which provide users with information on how to modify models in order to improve fit values (MacCallum, 1986; MacCallum et al., 1992; Sörbom, 1989). These modifications are based on statistical criteria and once conducted, shifts the research focus from confirmatory analysis to exploratory and data-driven analysis. Modification indices often show that model fit would improve if one or more residuals among indicator variables were allowed to correlate. At least some researchers will at this point, correlate the residuals among indicator variables. This practice is problematic for a variety of reasons.

## Problems with Allowing Correlated Errors

The first problem with allowing measurement errors to correlate in structural equation models based on post hoc modifications is that it allows researchers to achieve good fit statistics in spite of omitting relevant variables from their models (Cortina, 2002). As explained by Landis et al. (2009), “to the degree that two residuals correlate, there is evidence that there exists a cause of both of the variables to which the residuals are attached but that is not specified in the model” (p. 17). When indicator variables are systematically influenced by the same extraneous variable in addition to the specified latent variables they represent, the influence of the extraneous variable may be estimated through measurement error correlations without a specification of what the influence is (Fornell, 1983). As a result of the estimation of such correlations, the fit of the model improves, but our understanding of the phenomenon in question does not.

The second issue with allowing measurement errors to correlate in a post hoc fashion is that significant correlations are likely to be due to sampling error. Given that the number of off-diagonal elements of error covariance matrices can be very large, the probability of one or more such covariances being large simply because of sampling error is substantial. Several studies have shown that modified models often capitalize on the idiosyncratic features of sample data, and are likely to produce a departure from the true population model (Chou & Bentler, 1990; Green, Thompson, & Babyak, 1998; Green, Thompson, & Poirer, 1999; Lance, Conway, & Mulaik, 1988; MacCallum, 1986; MacCallum et al., 1992). To that end, Grant (1996) found that changing a hypothesized model to allow measurement errors to correlate based on specification search recommendations improved model fit in an initial sample, but failed to hold in cross-validation samples.

A third problem with allowing measurement errors to correlate is that this practice may bias parameter estimates of both the measurement and structural model (Tomarken & Waller, 2003). For example, Gerbing and Anderson (1984) argue that even when correlated measurement errors do not significantly alter parameter estimates of a measurement or structural model, they can still mask the underlying structure of modeled relationships.

## Previous Reviews

I will now turn our attention to previous reviews that have addressed the issue of error correlation in SEM. Three reviews have attempted to provide estimates of the extent to which published studies using SEM permit the errors among measurement errors to correlate. Unfortunately, these reviews report somewhat contradictory findings and sampled only a limited number of journals and studies.

The first review was conducted by Shah and Goldstein (2006), with a focus on this practice in management and operations journals. The authors examined the use of allowing correlated error and reviewed studies from *Management Science*, *Journal of Operations Management*, *Decision Science*, and *Journal of Production and Operations Management Society*, looking at all studies conducted from 1984 until 2003. While the practice of correlated errors was not the main focus of the study, it was included as an area of interest for the main topic, which was structural equation modeling practices. This review estimates that around 29% of articles testing CFA models freed the parameters among measurement error terms, while only 11% of articles testing strictly structural models engaged in this practice. Additionally, Shah and Goldstein state that only fifty percent of structural equation models that they reviewed provided justification for allowing measurement errors to correlate.

A second review was carried out by Cole, Ciesla, and Steiger (2007). The authors examined studies found in *Psychology Assessment*, *Journal of Counseling Psychology*, *Journal of Applied Psychology*, *Health Psychology* and *Journal of Personality and Social Psychology* in the year 2005. Across 75 studies, 21% of the articles explicitly allowed for error correlation, while an additional 5% almost certainly correlated errors, which was inferred from the calculation of degrees of freedom, while 5% of the articles were too vague to ascertain the presence of correlated errors. Therefore, according to the authors, anywhere from approximately 27% to 32% of published studies allowed measurement errors to correlate, an estimate quite different than that given by Shah and Goldstein (2006).

In addition to providing an estimate of allowing measurement errors to correlate when using SEM, Cole et al. ascertained what justifications were provided in the studies they reviewed for allowing errors to correlate. According to the authors, 71% of the justifications were driven by theory. That is, residuals were allowed to correlate when the measures were administered to the same informant. Twenty-nine percent of the justifications were driven empirically. That is, when residual correlations were allowed in order to generate a model that provided better fit to the data.

The authors also found that very few of the articles that allowed errors to correlate cross-validated the revised model with new data. This is undesirable, given that the practice of correlating errors can be seen as a capitalization on chance. Additionally, the focus of this study was an examination of the affects of not including design-driven correlated residuals in latent-variable covariance structure analysis. The stated conclusion derived by the authors was that failure to include certain correlated residuals can change the meaning of the extracted latent variables and generate potentially misleading results. Consequently, Cole et al. (2007) argue that after allowing measurement errors to correlate, authors should revise the description of the latent variables under question.

A final examination of the practice was examined by Landis et al., (2009). The authors examined 58 empirical articles derived from the *Journal of Applied Psychology*, *Journal of Management*, and *Personnel Psychology* that used structural equation modeling. According to the authors, 9 to 12% of published studies in these journals from 2002 to 2007 allowed measurement errors to correlate and approximately 70% of the researchers who engaged in this practice did so as a post hoc modification. Because the authors did not count articles as having correlated errors unless they were explicitly stated, it is likely that their estimates are an underestimation of the practice.

One reason for such contradictory findings is that these prior reviews is that studies covered came from a limited number of journals over brief spans of time. For instance, Landis et al. (2009) narrowed their review to studies published in *Personnel Psychology, Journal of Applied Psychology*, and *Journal of Management* from 2002 to 2007. Cole (2007) examined studies found in *Psychology Assessment*, *Journal of Counseling Psychology*, *Journal of Applied Psychology*, *Health Psychology* and *Journal of Personality and Social Psychology* from 2005. Finally, Shah and Goldstein (2006) reviewed studies from *Management Science*, *Journal of Operations Management*, *Decision Science, and Journal of Production and Operations Management Society*, although they did look at all studies conducted from 1984 until 2003. Clearly further investigation is needed in order to clarify the discrepancies found in these previous reviews by examining a broader range of journals over a larger span of time. Inow turn the discussion to factors that may influence the practice of error correlation.

### Model Complexity

Model complexity can be defined as the ability of a model to fit a diverse array of data patterns well by some established criterion of fit (Dunn, 2000; Myung, 2000; Pitt, Myung, & Zhang, 2002; Preacher, 2006). Previous research has suggested model complexity can be thought of as the average fit of a model to regions of data space, or the space containing all empirically obtainable data patterns relevant to a particular modeling domain (Preacher, 2006). Model complexity can be seen as the antithesis of parsimony. That is, all other things equal, as model complexity increases, parsimony is likely to decrease.

There are a number of factors that contribute to model complexity. Perhaps the most obvious is the effective number of free parameters in the model, which is defined as the number of freed parameters minus the number of functional constraints placed on otherwise free elements of the model. This difference is defined in SEM terminology as *q*. In structural equation modeling, all other things being equal, models with a higher q will be better able to fit data (Forster & Sober, 1994; Jeffreys, 1957; Wrinch & Jeffreys, 1921). This is because freeing model parameters reduces the number of dimensions in which observed data can differ from the data that would be implied by the hypothesized model (Mulaik, 2001; Mulaik, 2004). Q is inversely related (all else being equal) to degrees of freedom. That is, given a certain model, degrees of freedom represent the number of dimensions in which observed data can differ from the data that would be implied by the hypothesized model.

Another factor that contributes to model complexity is sample size. Sample size, like degrees of freedom is inherently related to the number of dimensions in which observed data can differ from the data that would be implied by the hypothesized model. Sample size is in turn connected with Chi-Square such that for virtually all models that are not perfectly clean fitting, Chi-Square increases as a function of sample size. Because virtually all of the major fit indices used today are derivatives of the Chi-Square in some way, sample size is inherently connected to fit index values in nearly all cases. It is therefore expected that the practice of error correlation will be related to model complexity via degrees of freedom and the sample size of the model in question.

One possible explanation for the allowance of correlated errors is that because extensive time, money, and effort goes into data collection and interpretation, researchers ultimately do not want to disregard data with poor fit and will instead attempt to save the data by improving fit through correlated errors (Hermida et al., 2010; Landis, Edwards, & Cortina, 2009). It has been conjectured that rather than abandon poor fitting data straight away, it might make more sense to modify the model so as to fit the data better (Sörbom, 1989). If correlated errors occur because of a researcher's desire to see good model fit for their data, then it stands to reason if a model already has acceptable fit, the likelihood of correlated errors is lessened. With respect to falsifiability, if power is associated with fit indices in covariance models such that more falsifiable studies are more likely to be rejected, and if researchers are motivated to correlate errors because of poor model fit, it stands to reason that falsifiability will be associated with the practice of error correlation such that the practice of error correlation will be positively related to the falsifiability of exact and close fit tests.

### Journal Impact and Influence

Another contribution that this study will offer is examination of how the practice of error correlation is influenced by the particular impact and influence of the journal in which the study was reported. Although the practice of correlated errors is widespread in psychology, many studies and reports have indeed exposed the problems associated with the practice (Brannick, 1995; Cliff, 1983; Gerbing & Anderson, 1984; Kaplan, 1989; Kaplan, 1990; Landis et al., 2009; MacCallum, 1986; MacCallum, Roznowski, & Waller, 1992; Shah & Goldstein, 2006; Steiger, 1990; Tomarken & Waller, 2003, Cortina, 2002). If the practice of correlated errors is indeed an undesirable practice, and journals of higher impact and influence are indeed of higher quality, it follows that the practice of error correlation should be less prevalent in higher quality journals. Therefore, it was hypothesized that the practice of error correlation will be related to journal impact and influence such that error correlation prevalence will be negatively correlated with journal impact and influence.

### Unique Contributions and Aims

While there have been attempts to capture the practice of correlating errors in structural equations modeling, this quantitative review seeks to provide additional contributions over and above those of the three previous studies. First, the three previous studies differ in their estimations of prevalence for the practice of allowing correlated errors in SEM. This is most likely because each study used a small sample of journals across a narrow range of disciplines, and on most occasions, analyzed studies over different time periods. The current study seeks to remedy this problem by examining a much wider array of journals over a consistent time period. More specifically, this review will examine a wide range of journals over a ten year time period from 1997 to 2007. By casting a wider net of journals and disciplines over a longer and consistent time period, this quantitative review will provide a more unifying and comprehensive examination of the degree to which correlated errors are practiced. A question of interest for this study is the degree and form in which the practice occurs in specific subdisciplines in psychology and disciplines other than psychology, such as management and educational research. Examination of this question allows us to determine if some fields of study allow researchers to engage in this practice to a greater extent than others, and thus discuss the implications this has on conclusions garnered from SEM research in specific sub domains.

A second major contribution of this study is a fuller and more in-depth examination of the justifications researchers provide for allowing measurement errors to correlate. Certainly, improved model fit is a common justification, but it is possible that other explanations exist. For example, it is not uncommon for studies to allow for correlated errors in initial model testing when the research design is longitudinal and the errors that are allowed to covary are the same indicators at different time periods. It is also possible for researchers to hypothesize a priori that errors will be correlated, based on the nature of study variables. For example, many studies allow errors to correlate when the variables have shared components. Kenney and Judd (1984) suggested using all possible cross-products of latent variable indicators as indicators of a latent product to be used for testing multiplicative structural equation models. Some of these cross-products will share components, so it is almost certain that their errors will correlate. Ultimately, these two reasons for allowing errors to correlate are part of the design, and are not necessarily related to sampling error or omitted variables issues. Thus, this study will help determine if the majority of studies which allow measurement errors to correlate are doing so for theoretically justifiable reasons, such as longitudinal research, or for unjustifiable reasons, such as improvement of model fit.

In this study I also attempted to examine the less appropriate justifications for allowing errors to correlate, specifically when they are made ex-post facto. I determined the rate of these justifications and if any articles in particular are being citied as support for this being a valid practice. An attempt will be made to match what was ascertained and stated by the researchers in the original article and what was ascertained and stated by the researches in the correlated errors study to examine if there are any disconnects between the two articles. It is my expectation that for some of the correlated errors studies, where the correlation occurred ex-post facto, there will be some miscommunication between cited articles and application of the ideas in those articles. By uncovering a source of miscommunication, I can begin to unlock the antecedents to this practice and will be in a better position to offer recommendations for practice.

Finally, no study to date has examined the idea of model complexity being related to the practice of correlating errors. One of the most parsimonious explanations for why researchers allow correlate errors in structural equation modeling is that they do not wish to abandon data that has poor fit; therefore they will correlate errors unjustly for the purpose of having a better fitting model. If this is true, it stands to reason that antecedents to poorer fitting data might be relevant to the practice of correlating errors. One such antecedent is model complexity. All other things being equal, a more complex model will generate better overall model fit. Therefore, it seems more complex models will be less likely to engender the practice of correlating errors.

## Method: Literature Search

The goal of this study was to review the practice of allowing correlated measurement errors in a variety of disciplines. Thus, studies from various fields of study including Psychology, Education, Business, and Sociology were included. Consequently, PSYCINFO, ProQuest, ERIC, AB-INFORM, were used to collect empirical articles. Article searches were limited to years 1997 through 2007 in order to represent modern methodological practices. In all databases, keywords covariance models, structural equation modeling, and SEM were utilized in order to identify all articles that used structural equation modeling.

## Method: Summary of Meta-Analytic Dataset

From searches of the aforementioned databases, 315 useable articles that that allowed measurement errors among indicator variables to correlate were identified. I excluded literature reviews, methodological papers, and papers that created models with simulated data. The 315 articles were coded by two different coding pairs, with each all coders possessing advanced degrees related to Psychology. Table 1 illustrates interrater agreement statistics.

**Table 1.** Interrater Agreement Statistics

| Meta-Analytic Category | Type of Agreement | Agreement Value |
|---|---|---|
| Sample Size | Intraclass Correlation Coefficient | 0.98 |
| Degrees of Freedom | Intraclass Correlation Coefficient | 0.82 |
| Journal Quality | Intraclass Correlation Coefficient | 0.99 |
| Fit Index Values | Intraclass Correlation Coefficient | 0.98 |
| Correlated Errors (yes/no) | Cohen’s Kappa | 0.97 |
| Model Type (measurement/structural) | Cohen’s Kappa | 0.95 |
| Rationale for Correlating Errors | Cohen’s Kappa | 0.88 |

### Method: Inclusion Criteria

In order to begin coding articles, it first had to be determined which studies allowed measurement errors to correlate. The articles were searched for any combination (or variant) of the following terms: correlated, freed, errors, and residuals. The models themselves were also visually inspected for measurement error pathways. It is possible and probable that some published manuscripts neither mentioned nor depicted the measurement error pathway even if the author did free the path. Unfortunately, in the absence of those explicit text or diagram indicators there is no feasible method for including those manuscripts that are consistent across all articles.

### Method: Treatment of Multiple Reported Models

When multiple tested models were reported in a single manuscript additional decisions were made in coding. The inclusion of correlated measurement error pathways is usually done based upon the modification indices from the results of a tested theoretical model. The model with the included correlated measurement error pathways is referred to here as the modified model. When the coders were able to determine which reported model contained correlated measurement errors the model tested just previous was included as the theoretical model. The two models were included in the coding as the theoretical model and the modified model.

### Method: Coding Procedures

Once all studies allowing the measurement error pathways were identified, relevant moderators were also coded. Moderator selection was based upon there potential influence on model statistics, model complexity, authors' decisions, and reporting standards. Those four classifications encompass both the impact of correlated measurement errors on model parameters and the conditions under which researchers' practice SEM. Model statistics serve as the measure of impact from the correlated measurement error pathway inclusion. Model complexity is a potential reason for the correlation of measurement errors, and also has the potential to directly affect model statistics. Each of the classifications is further detailed below.

### Method: Model Statistics

When theoretical and modified models were reported, both sets of fit statistics, degrees of freedom, and statistical significance were coded. Model fit statistics were differentially reported but every fit statistic value reported was coded. If the pathway that correlated errors was reported, that value was also recorded in terms of the correlational relationship. When more than one such pathway was present an average correlation was recorded. In addition, the model degrees of freedom and statistical significance were coded. The three codes for model significance included if both the theoretical and modified models were non-significant, if statistical significance changed to non-significance, and if the model remained significant after allowing correlated measurement errors.

### Method: Model Complexity

Model complexity moderators include the model type, the number of manifest items, and the number of latent constructs. The presence or absence of item parcels and the range of items per parcel were coded to ensure accurate reflection of model complexity. The models where coded as either a measurement model or structural model.

### Method: Author Decisions

The researchers' motives and justifications for allowing measurement errors to correlate were coded. A distinction was made between a priori and longitudinal justifications. Models that coded the measurement errors post-hoc were categorized into 3 groups: as related to model content if stated by the authors; methodological if the authors stated that parameters between measurement errors were freed because of the same measure(s) being used over time or if they were specified due to estimating a model with a moderator and allowing the moderator indicator residuals to correlate with the component indicators residuals (but not if the component indicators were allowed to correlate); and data driven if the researcher specified that measurement errors were allowed to correlate because they improved model fit. Any citations given as a justification were also coded. These cited references were subsequently reviewed by the coders and a judgment was made if the cited reference matches the reported justification.

### Method: Reporting Standards

Authors present and report model parameters in conformity to their research field. As the field and journal characteristics are potential influences on the inclusion or exclusion of measurement error pathways, they were coded. It was hypothesized that higher impact journals would be stricter in allowing the practice of correlating measurement errors. The journal impact and value are based upon the year the article was published (Bergstrom, 2007).

### Results: Point Estimation

One aim of this research was to establish a point estimate for the percentage of studies that correlated errors in the psychological literature. To examine this issue, a random selection of 985 studies that used some form of structural equation modeling was obtained. Next, I calculated the number of studies that correlated errors. The results indicated that within these studies, 315 studies correlated errors out of a possible 985 studies. The percentage of studies that correlate errors in the psychological literature is thus best estimated at 32%. The average correlation between correlated errors was 0.31.

### Results: Rationales Given

Of critical interest to this study was the rationale researchers gave for error correlation. Rationales were classified under five categories: a-priori, longitudinal designs, post-hoc construct theory, post-hoc method theory, and modification indices. Listing of the percentages by rationale appears in Table 2. The most common rationale was correlating errors as a result of modification indices specified after running the initial model. It should be noted that only 25% of the models were correlated for a-priori reasons, or because the data in question was longitudinal. Additionally of interest were the particular citations given by researchers. Unfortunately, most researchers did not cite a specific article in support of error correlation. Specifically, only 15% of studies gave a direct citation in support of the choice to correlate errors. Out of the 15% of articles that gave a specific citation, the vast majority of those citations were related to post-hoc construct theories, with the rationale equating to the correlation of measurement errors based on the fact that a different study had found the variables to be significantly associated with one another.

**Table 2.** Rationale for Error Correlation

| Rationale | Percentage of Sample |
|---|---|
| Modification Indices | 37% |
| Post-Hoc Construct Theory | 24% |
| Longitudinal Data | 18% |
| Post-Hoc Method Theory | 14% |
| A-Priori Theory | 7% |

### Results: Model Complexity

Another aim of this research is to examine the model complexity of covariance models that correlated errors vs. the model complexity of covariance models that did not correlate errors. Indeed, there was a statistically significant difference in the degrees of freedom of models in studies that correlated errors (M = 73.25, SD = 8.23), and those that did not (M = 69.02, SD = 6.34), $t(920) = 12.36$, $p < .05$, d =.58. Similarly, there was a statistically significant difference in the sample size of models in studies that correlated errors (M = 314.18, SD = 64.23), and those that did not (M = 279.55, SD = 74.21), $t(920) = 10.71$, $p < .05$, d =.50. As expected, models that correlated errors were significantly more complex, and were significantly more falsifiable than models that did not correlate errors.

## Results: Journal Impact

Another aim of this research was to establish the relationship between journal impact and the practice of error correlation. To examine this issue, the journal impact for all studies was collected. Next, each study was assigned a binary code to assess if they correlated errors (0 = no 1 = yes). Logistic regression analysis was conducted with the dichotomous variable of error correlation as the dependent variable, and journal impact as the independent variable. The results indicated that indeed, journal impact was associated with the practice of error correlation such that as journal impact increased, the probability of error correlation decreased, (odds ratio [OR] = 0.61, 95 % confidence interval [CI] = 0.55-0.68, *p*<.01).

## Results: Fit Index Improvement

Another important aim of this research is to establish the degree to which fit indices are impacted by the practice of error correlation. To examine this issue, two sets of fit indices were evaluated: those calculated by the researcher before error correlation and those calculated by the researcher after error correlation. I examined this issue across studies where error correlation was conducted for invalid reasons (i.e. not a-priori or longitudinal). Table 3 displays the change in fit across the examined fit statistics. In general, the change in fit as a result of error correlation resulted in a .02-.03 betterment in fit across indices. This difference amounts to approximately half of the difference between qualitatively different assessments of fit, such as a value of.08 on the RMSEA fit statistic corresponding to “moderate” model fit, and a value of.05 (difference of.03), or such as a value of.90 on the CFI fit statistic corresponding to “reasonable” fit, and a value of.95 (difference of.05) corresponding to “close” fit (Hu & Bentler, 1999).

Because of the possibility of researchers correlating models to move through the unofficial rules of thumb with respect to fit, the percentage of studies that correlated errors to pass through “good” and “excellent” thresholds of model fit was examined. Specifically, I examined the percentage of models that did not meet a fit index value associated with the “rules of thumb”, and the percentage of models that passed through that threshold after the correlated or errors. I examined the RMSEA and CFI fit indices, with the rules of thumb equating to RMSEA values of .05 and .08 for good and moderate fit, respectively (Brown & Cudeck, 1992), and CFI values of .90 and .95 for reasonable and close fit, respectively (Hu & Bentler, 1999).

Out of the all the models that a) presented both pre and post correlation RMSEA fit values, and b) failed to pass the .08 threshold for RMSEA prior to the correlation of errors, 79 percent of models passed the .08 threshold for model fit *after* the correlation of errors. Similarly, out of the all the models that a) presented both pre and post correlation RMSEA fit values and b) failed to pass the .05 threshold for RMSEA prior to the correlation of errors, 34 percent of models passed the.05 threshold for model fit *after* the correlation of errors.

Out of the all the models that a) presented both pre and post correlation CFI fit values, and b) failed to pass the .90 threshold for CFI prior to the correlation of errors, 78 percent of models passed the .90 threshold for model fit *after* the correlation of errors. Similarly, out of the all the models that a) presented both pre and post correlation CFI fit values and b) failed to pass the .95 threshold for RMSEA prior to the correlation of errors, 48 percent of models passed the .05 threshold for model fit *after* the correlation of errors.

**Table 3.** Fit Statistic Change as a Result of Error Correlation

| Model | DF | $\chi^2$ | RMSEA | CFI | GFI | NNFI |
|---|---|---|---|---|---|---|
| Before Error Correlation | 75 | 395.03 | .09 | .92 | .92 | .90 |
| After Error Correlation | 73 | 328.73 | .07 | .95 | .94 | .93 |
| Difference | 2 | 66.30 | .02 | .03 | .02 | .03 |

## Discussion

With few exceptions, there is no theoretical defensible reason for the practice of error correlation. As found in the current study, nearly one-third of all studies that use structural equation modeling engage in the practice of measurement error correlation, and of those that do, most researchers engage in this practice for invalid reasons. Consequently, the current quantitative review has shown the severity of this problem across various fields of study, and highlights the need for SEM users, journal reviewers, and researchers to be wary of misspecified models found in the literature as well as the need for more stringent guidelines for reviewing and accepting SEM manuscripts. If the errors that are correlated are random in nature, then correlating errors is taking advantage of random chance, thereby moving the researcher from confirmatory model testing to exploratory model testing. Alternatively, if the errors that are correlated are the result of omitted variables, it is imperative to identify the missing variables, collect data from a second sample, and test the hypotheses that the omitted variable accounted for the correlated error (Cortina, 2002). For both of those reasons, if error correlation is seen in a poor-fitting model's modification indices and is not a result of the study's design, then the best possible course of action is most likely to form a hypothesis about the reason for the errors being correlated, such as hypothesizing what variable was omitted in the model, and then test the new model with new and independent data (Hayduk & Glaser, 2000).

### Summary of Findings

Researchers most frequently engaged in the practice of allowing measurement errors to correlate because of the increases to be gained in model fit shown by the modification indices. Apart from modification indices, researchers most frequently correlated errors because the two variables in question were cited as being associated with one another in a different study. This justification for allowing measurement errors to correlate post hoc is invalid and atheoretical for many reasons.

First, if a researcher can provide a strong justification for allowing correlated errors as a result of a known omitted variable, it is reasonable to wonder why the parameter was not represented in the original model (MacCallum et al., 1992). Second, performing such modifications is tantamount to rewarding poor scale construction and/or model development. Third, there remains a potential disconnect between the omitted variable claimed to have caused the error correlation in the model, and the omitted variable that actually caused the error correlation in nature. In psychology, it has been said that "everything correlates with everything else", given rise to what David Lykken (1968) called "the crud factor" (Meehl, 1997). If indeed variables have associations with numerous other variables, it seems unlikely that the researcher in question could accurately ascertain the exact nature of what caused the error correlation in that particular model without conducting a follow up study and cross-validating the model.

The current study also demonstrated how error correlation improves fit in a non-trivial way. The improvement in model fit was on average approximately half of the distance between models that are judged to be adequate according to the rules of thumb, and models that are judged to be good. Related to that point, approximately three-quarters of models that prior to error correlation did not reach the minimum threshold for fit, according to the rules of thumb, reached minimum fit after error correlation. This is problematic, and suggests that researchers correlate errors to pass through the unofficial rules of thumb SEM practitioners have set regarding model fit and model testing in general. This could be because models that do not pass through these rules of thumb are less likely to be published after going through the review process and researchers are thus rewarded for using modification indices to obtain value on the sunk costs they put into data collection. Therefore, it seems that researchers are incentivized to correlate errors, when faced with a model that is within range of passing the thresholds associated with the rules of thumb.

### Recommendations and Future Research

Now that the extent of the problem of allowing measurement errors to correlate in structural equation models and the reasons for this problem are known, some potential remedies can be offered.

One remedy that is offered is for reviewers and referees to abandon qualitative rules of thumb descriptions to model fit based on unofficial rules of thumb. I make this recommendation for several reasons. First, attaching qualitative labels to quantitative factors (such as model fit) causes loss of information in much the same way as dichotomization of continuous variables will cause a loss of information. To give an extreme example, RMSEA values of.079 and.081 fit an entire covariance model to virtually the same degree (all else being equal), but would have different qualitative labels of model fit, if the rules of thumb are followed. It is also known that optimal cutoff criteria are heavily dependent on specific aspects of the model apart from fit (Marsh et. al, 2004; Nye & Drasgow, 2011). Specifically, simulation research has shown that the optimal cutoff criteria are actually dependent on numerous features of the model, including the estimation method used, the sample size used, the number of free parameters, and the degree to which assumptions of multivariate normality are met or not met (Hu & Bentler, 1999, Marsh et. al, 2004; Tomarken & Waller, 2005). Because of this, traditional fit indices, models evaluation of model fit should be considered with respect to the model's specific characteristics. Cohen (1988) cautioned against strict interpretations and uses of rules of thumb with respect to effect size, arguing for a more nuanced and thoughtful approach. This recommendation is also applicable to interpretation of fit indices in SEM and would hopefully eliminate a possible incentive researchers have for correlating errors inappropriately, without harming proper interpretation of fit indices.

An additional consideration regarding fit indices relates to a second issue: the use of approximate model fit indices to evaluate model fit. Some researchers have argued that Chi-Square is the only acceptable fit index to use in evaluating structural equation models (Barrett, 2007; Hayduk, Cummings, Boadu, Pazderka-Robinson, & Boulianne, 2007) The argument for the sole use of the Chi-Square in evaluating models is centered on the following points: 1) there are no single thresholds for GOF indices that can be applied under all possible measurement and data conditions, 2) GOF indices allows for researchers to avoid careful model specification and examination, 3) GOF indices can allow mediocre models to make it through the peer-review process, 4) the potential for the seriousness of casual misspecification to be uncorrelated with GOF indices values, and 5) Chi-Square does a better job at detecting model misspecification than does any other fit index.

Readers who are interested in examining this issue at length can examine the above issues and counterarguments in the special issue of the *Personality and Individual Differences* journal that summarizes this debate. I would simply point out that to the degree approximate model fit indices are inappropriate; is the degree that use of GOF indices provides little *benefit* for scientific advancement, considering the *cost* of prompting researchers into inappropriate error correlation.

The second recommendation is for researchers to engage more seriously in cross-validation and replication of models. If a researcher believes they understand the cause for modification indices turning up significant pathways between measurement errors, the researcher should collect data from a new sample, with all variables included, thus confirming the model uncovered by exploratory analysis.

The third and final recommendation is for psychology to engage in more serious enforcement of quantitative sections of studies that use SEM and also to engage in more serious quantitative education with respect to SEM. A main theme from this quantitative review is that inappropriate statistical practices most likely stem from either misunderstanding of statistical issues, or through the mechanism of individual self-interest (or both). If reviewers and editors insist on researchers not improperly correlating errors, then the self-interest component of the researcher will be oriented towards not correlating errors in inappropriate situations. Additionally, if more serious quantitative education with respect to SEM is undertaken, it seems likely that researchers who are concerned with quantitative methods will not correlate errors improperly.

## Limitations

This study contains some limitations. First, this quantitative review only contained articles that made it through the publication process. While this was intentional, as the focus was specifically on published works, it could be the case that the practice of inappropriate error correlation differs from published to non-published studies. For example, it could be the case that researchers who refuse to inappropriately correlate errors are underrepresented in published articles, thus altering the percentage of studies that correlate errors. Therefore, it is important to not generalize the findings in this quantitative review to non-published studies. Second, there are cases where refusing to correlate measurement errors might make solid model fit an especially conservative undertaking. For example, residual correlation has been recommended in multiple mediation models because covariances among the mediators are unlikely to be completely explained by their mutual dependence on the independent variable (Preacher & Hayes, 2008). While I take the position that residual and measurement error correlation is to be avoided in this type of case, it should also be recognized that by doing so, model fit will possibly be slanted in a conservative manner to some degree, and there remains a possibility for model misspecification as a result of not permitting residual covariances to vary, especially if the those covariances are significant in magnitude.

## Conclusions

This study provides strong evidence for the importance of understanding the practice of inappropriate error correlation in structural equation modeling. Methodological issues have been emphasized in structural equation modeling in scientific research and education in quantitative methods. An additional issue to emphasize in these venues should be the abolishment of inappropriate error correlation practices.

## References

[1] Markon, K.E., & Kruger, R.F. "*An empirical comparison of information-theoretic selection criteria for multivariate behavior genetic models*", Behavior Genetics, 34. 593-610. 2004.

[2] MacCallum, R.C., & Austin, J.T. "*Applications of structural equation modeling in psychological research*", Annual Review of Psychology, 51.201-226.2000.

[3] Brannick, M.T. "*Critical comments on applying covariance structure modeling*", Journal of Organizational Behavior, 16. 201-213.1995.

[4] Cliff, N. "*Some cautions concerning the application of causal modeling methods*", Multivariate Behavioral Research, 18.115-126.1983.

[5] Cortina, J. "*Big things have small beginnings: An assortment of 'minor' methodological misunderstandings*", Journal of Management, 28. 339-262. 2002.

[6] Gerbing, D.W., & Anderson, J.C. "*On the meaning of within-factor correlated measurement errors*", Journal of Consumer Research, 11. 572-580. 1984.

[7] Kaplan, D. "*Model modification in covariance structure analysis: Application of the expected parameter change statistic*", Multivariate Behavioral Research, 24. 41-57. 1989.

[8] Kaplan, D. "*Evaluating and modifying covariance structure models: A review and recommendation*", Multivariate Behavioral Research, 24. 137-155. 1990.

[9] MacCallum, R.C. "*Specification searches in covariance structure modeling*", Psychological Bulletin, 100. 107-120. 1986.

[10] MacCallum, R.C., Roznowski, M., & Necowitz, L.B. "*Model modifications in covariance structure analysis: The problem of capitalization on chance*", Psychological Bulletin, 111. 490-504. 1992.

[11] Shah, R., & Goldstein, S.M. "*Use of structural equation modeling in operations management research: Looking back and forward*", Journal of Operations Management, 24. 148-169. 2006.

[12] Steiger, J.H. "*Structural model evaluation and modification: An interval estimation approach*", Multivariate Behavioral Research, 25. 173-180. 1990.

[13] Tomarken, A.J., & Waller, N.G. "*Potential problems with "well fitting" models*", Journal of Abnormal Psychology, 112. 578-598. 2003.

[14] Landis, R., Edwards, B. D., & Cortina, J. "*Correlated residuals among items in the estimation of measurement models*". In C. E. Lance & R. J. Vandenberg (Eds.). Statistical and methodological myths and urban legends: Doctrine, verity, and fable in the organizational and social sciences (pp. 195-214). New York: Routledge. 2009.

[15] Anderson, J.C., & Gerbing, D.W. "*The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis*", Psychometrika, 49. 155-173. 1984.

[16] Jöreskog, K.G. "*A general approach to confirmatory maximum likelihood factor analysis*", Psychometrika, 34. 183-202. 1969.

[17] Maruyama, G. "*Basics of Structural Equation Modeling*". Thousand Oaks. Sage. 1998.

[18] Sörbom, D. "*Model modification*", Psychometrika, 54. 371-384. 1989.

[19] Fornell, C. "*Issues in the application of covariance structure analysis: A comment*", Journal of Consumer Research, 9. 443-448. 1983.

[20] Chou, C. P., & Bentler, P. M. "*Model modification in covariance structure modeling: A comparison among the likelihood ratio, Lagrange Multiplier, and Wald tests*", Multivariate Behavioral Research, 25. 115–136. 1990.

[21] Green, S. B., Thompson, M. S., & Babyak, M. A. "*A Monte Carlo investigation of methods for controlling Type I errors with specification searches in structural equation modeling*", Multivariate Behavioral Research, 33. 365–383. 1998.

[22] Green, S. B., Thompson, M. S., & Poirer, J. "*Exploratory analyses to improve model fit: Errors due to misspecification and a strategy to reduce their occurrence*", Structural Equation Modeling, 6. 113–126. 1999.

[23] Lance, C.E., Cornwell, J.M., & Mulaik, S.A. "*Limited information parameter estimates for latent or mixed manifest and latent variable models*", Multivariate Behavioral Research, 23. 171-187. 1988.

[24] Grant, L. D. "*A comprehensive examination of the latent structure of job performance*", Ph.D. dissertation, North Carolina State University. 1996.

[25] Cole, D., Ciesla, J., & Steiger, J. "*The insidious effects of failing to include design-driven correlated residuals in latent-variable covariance structure analysis*", Psychological Methods, 12, 381-398. 2007.

[26] Dunn, J.C. "*Model Complexity: The fit to random data reconsidered*", Psychological Research, 63. 174-182. 2000.

[27] Myung, I. J. "*The importance of complexity in model selection*", Journal of Mathematical Psychology, 44. 190–204. 2000.

[28] Pitt, M.A., Myung, I. J., & Zhang, S. "*Toward a method of selecting among computational models of cognition*", Psychological Review, 109. 472-491. 2002.

[29] Preacher, K.J. "*Quantifying Parsimony in Structural Equation Modeling*", Multivariate Behavioral Research, 41. 227-259. 2006.

[30] Forster, M.R., & Sober, E. "*How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions*", British Journal for the Philosophy of Science, 45. 1-35. 1994.

[31] Jeffreys, H. "*Scientific inference*", Cambridge University Press, Cambridge. 1957.

[32] Wrinch, D., & Jeffreys, H. "*On certain fundamental principles of scientific inquiry*", The London Edinburgh and Dublin Philosophical Magazine and Journal of Science, 42. 339-369. 1921.

[33] Mulaik, S.A. "*The Curve-fitting problem: An objectivist view*", Philosophy of Science, 68. pp 218-241. 2001.

[34] Mulaik, S.A. "*Objectivity in science and structural equation modeling*". In David Kaplan (Ed.), The Sage Handbook of Quantitative Methodology for the Social Sciences. Thousand Oaks. Sage. 2004.

[35] Hermida, R., Conjar, E.A., Najab, J.A., Kaplan, S.A., & Cortina, J.M. "*On the Practice of Allowing Correlated Residuals in Structural Equation Models*". Unpublished Manuscript, Department of Psychology, George Mason University, Fairfax, Virginia, United States. 2010.

[36] Kenney, D., & Judd, C.M. "*Estimating the Non-linear and Interactive Effects of Latent Variables*", Psychological Bulletin, 96. 201-210. 1984.

[37] Bergstrom, C. Eigenfactor.org: *Ranking and Mapping Scientific Knowledge,* from http://www.eigenfactor.org/. 2007.

[38] Browne, M.W., & Cudeck, R. "*Alternative ways of assessing model fit*", Sociological Methods and Research, 21. pp. 230-258. 1992.

[39] Hu, L., & Bentler, P. "*Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives*", Structural Equation Modeling, 6. 1-55. 1999.

[40] Hayduk, L.,&Glaser, D. N. "*Jiving the four-step, waltzing around factor analysis, and other serious fun*". Structural Equation Modeling, 7. pp. 1-35. 2000.

[41] Lykken, D.T. "*Statistical significance in psychological research*", Psychological Bulletin, 70. pp 151-159. 1968.

[42] Meehl, P. E. "*The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions*". In L. L. Harlow, S. A. Mulaik, & J.H. Steiger (Eds.), What if there were no significance tests?, Lawrence Erlbaum Associated, Mahwah, pp. 393-425. 1997.

[43] Marsh, H., Hau, K., & Wen, Z. "*In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings*", Structural Equation Modeling, 11. 320-341. 2004.

[44] Nye, C. D., & Drasgow, F. "*Assessing goodness of fit: Simple rules of thumb simply do not work*", Organizational Research Methods, 14. pp. 548-570. 2011.

[45] Tomarken, A., & Waller, N. "*Structural Equation Modeling: Strengths, Limitations, and Misconceptions*", Annual Review of Clinical Psychology, 1. 31-65. 2005.

[46] Cohen, J. "*Statistical power analysis for the behavioral sciences*", Lawerence Erlbaum Associates, Hillsdale, New Jersey. 1988.

[47] Barrett, P. "*Structural equation modeling: Adjudging model fit*", Personality and Individual Differences, 42. pp. 815-824. 2007.

[48] Hayduk, L., Cummings, G., Boadu, K., Pazderka-Robinson, H., & Boulianne, S. "*Testing! testing! One, two, three-Testing the theory in structural equation models!*", Personality and Individual Differences, 42. pp. 841–850. 2007.

[49] Preacher, K.J., & Hayes, A.F. "*Asymptotic and resampling strategies for assessing and comparing indirect effect in multiple mediator models*", Behavior Research Methods, 40. pp. 879-891. 2008.