The Impact of Unmodeled Error Covariance on Measurement Models in Structural Equation Modeling

Fraser D Bocell

A Dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2015

Supervisory Committee:

Elizabeth Sanders
Robert Abbott
Min Li
Deborah McCutchen
Dagmar Amtmann

Program Authorized to Offer Degree:

College of Education

University of Washington

Abstract

The Impact of Unmodeled Error Covariance on Measurement Models in Structural Equation Modeling

Fraser Bocell

Chair of the Supervisory Committee:
Professor Elizabeth Sanders, Ph.D.
Educational Psychology

Subject responses to observed variables are affected by both the underlying construct(s) of interest as well as the observed variables' measurement error. In practice, method effects can cause some observed variables to share error, violating the assumption of linear independence if the error covariance is not directly modeled. The current study evaluated the impact of failing to account for error covariance under varied, real-world conditions. Data from a 2-factor, 4-indicator per factor confirmatory factor analysis (CFA) model were simulated to have one correlated error term between two items that load onto the first factor. Levels of within-factor error correlation, sample sizes, loading magnitudes, and factor-factor correlations were varied, with *N* = 100,000 simulations per condition. All conditions were analyzed using a 2-factor model that assumed no error covariance, leaving the assumption of linear independence intact. The results were evaluated for the effect of the misspecification of the correlated error term on estimate bias as well as 95% confidence interval coverage. Overall, the findings showed that the failure to account for correlated error terms caused bias in parameter estimates that was exacerbated by the magnitude of the covariance between error terms. Greater shared error variance between items led to larger bias and poorer confidence interval coverage, while larger factor loadings decreased the amount of bias and improved confidence interval coverage. Relative bias in the factor loadings was not affected by sample size, though confidence interval coverage worsened with larger samples. Discussion of results and recommendations for applied analysts are given.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Acknowledgements

The author would like to thank all of the members of his committee for the support they have given him throughout this process. Especially, Bob, Min and his advisor Liz, who have supported, guided, and taught him throughout his years at UW. He would also like to thank his previous advisor, Cathy Taylor for her support and guidance. Thanks is also due to his friends and family for their always generous support.

The data for the applied analysis was generously provided by Denise Wilson, Ph.D. and the REESE program. The author's work on this program served as the inspiration for this study. The author would also like to than Diane Carlson-Jones, Ph.D. for her help in obtaining permission to use the data.

The author would also like to gratefully acknowledge the National Science Foundation for their support of this work under the REESE program (grant numbers DRL-0909817, 0910143, 0909659, 0909900, and 0909850). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Chapter I: Introduction and Statement of Problem

Ways of Measuring

Measurement is a fundamental ingredient of all research activities. However, it is important that we distinguish between two kinds of measures from the outset: those that are *observable* and those that are *representational* (Kline, 2010). Observable measures are those that have a one-to-one correspondence between the numerical value collected and the quantity of what is being measured (e.g., blood pressure reading in millimeters of mercury, or amount of electrical activity in a brain region in millivolts, etc.); for observable measures, the number we collect is in units that are directly tied to what was measured. These kinds of measures are falsifiable; that is, measurement instruments can be readily checked for accuracy and precision.

Representational measures, on the other hand, are those in which a theoretical idea ("latent" construct or measure) that cannot directly be observed is instead operationally defined in order to construct a measurement procedure that should represent aspects of the original idea – the numerical quantity obtained from these instruments does not necessarily relate directly back to the units of the larger, theoretical idea (e.g., a numerical score of 5 does not translate directly to "5 units" of the theoretical idea). It is difficult to conceive of any survey, test, or questionnaire that is not a representational measure.

When we use representational measurement instruments, we must infer that the observed data are *reflective* of a quantity (or quality) of the larger, unobserved, latent construct. As an example, a child's observed score on an arithmetic test constructed by a certain publisher may be thought to mirror a portion of the child's *true* math skill or ability – a

skill or ability which certainly encompasses much more than simple arithmetic operations. If the test is accurately assessing a portion of true math skill, then we would naturally conclude that a child with a high level of math skill would likely have a high score on this test, and similarly, that a child with a lower level of math skill would have a lower score on the same test. Importantly, scores from this hypothetical arithmetic test could then be used in a number of ways in education research and practice, including: as educational feedback for the child and his/her parents (e.g., determining whether the child should receive special or gifted services), as a policy indicator about a teacher, school, district, or state (e.g., children in School X scored very low on this test and therefore School X may need to change its math curriculum), as part of an educational research study (e.g., children's scores on this test may have relationships with other skills or attitudes, or perhaps particular types of math curricula and instruction) or as evidence of the test's validity (e.g., construct validity, discriminant validity, etc.).

Despite the usefulness and practicality of representational measures, particularly in the absence of directly observable measures, their use is not without drawbacks. First, as already noted, we cannot conclude that the observed score on any representational measure is directly equal to the number of units of the true latent construct. Second and more importantly, unlike observed measures, there is difficulty in evaluating the accuracy and precision of the instrument in measuring the true latent construct. While there are numerous methods for evaluating measure construct validity (cf. Messick, 1990), latent variable modeling (also known as "factor analyses" and "structural equation modeling") is one of the primary analytical tools researchers can employ for assessing items on a single measure simultaneously, and by extension, multiple measures simultaneously (e.g., a theoretical multi-measure system). Since

the advent of relatively easy-to-use computer software in the 1990s, latent variable modeling has become widely employed across educational research today (e.g., Bruning, Dempsey, Kauffman, McKim, & Zumbrunn, 2013; Marsh et al., 2013; Purpura & Lonigan, 2013).

Latent Variable Modeling

Unlike other forms of establishing validity evidence, latent variable statistical models allow us to formally test the relationships between the items of an instrument (or variables among several instrument(s) and the construct(s) those variables are meant to measure. With these models, researchers can estimate the relative strength of these relationships ("measurement model"), the relationships among constructs if there is more than one being measured ("structural model"), and importantly, overall model fit. These estimates in turn allow researchers to make decisions about how well the items/variables represent the construct(s), whether some items/variables should be dropped from the system, and whether additional latent factors should be added to the system. Based on analysis results, scales can then be created for use in the field (e.g., applied educational research and practice).

These latent models are broadly termed structural equation models (SEMs), with specific kinds of SEMs characterized by the number of relationships and factors estimated, as well as the measurement scales of both the observed and hypothesized latent variables. For example, an exploratory factor analysis (EFA) is an SEM that estimates a relationship among all items/variables with all factors; a confirmatory factor analysis (CFA), on the other hand, is one in which the item-factor relationships are constrained (based on empirical or theoretical frameworks) and therefore represents a simplification of the EFA. Another way to envision the difference is through the concept of a priori and a posteriori latent variables (Bollen, 2002). EFA

relates to a posteriori latent variables, where the latent construct is defined after the data has been analyzed. EFA is used to locate and define the latent variable. In contrast, CFA seeks to confirm the presence of a theoretical latent variable defined before data analysis, an a priori latent variable. An EFA and CFA can then be compared for overall model fit – the simplest model (i.e., the one that estimates the fewest relationship parameters), but that fits the data sufficiently well, is considered the preferred model.

Confirmatory Factor Analysis

Prevalence of CFA in applied research. Confirmatory Factor Analysis (CFA) is one type of structural equation model, based on the analysis of covariance structures. CFA is used to analyze the underlying structure of a given set of observed items from a test, evaluation or survey. As Bollen wrote, "The primary goal [of CFA] is to explain the covariances or correlations between many observed variables by means of relatively few underlying latent variables"(1989, p. 226). It is in the field of measurement and assessment that CFA has become a staple statistical procedure, beginning with work by Spearman (1904) in explaining subjects' test performance using a single "General Intelligence" factor.

Before the use of the CFA became common, Cole (1987) was aware of the potential importance of CFA methods in test validation studies, and he advocated for expanding their use throughout the field. More recently, Kline noted the importance of CFA to many researchers, "especially those who conduct assessment-related studies" (2010, p. 230). This sentiment is also shared by Taylor (2013) in her book *Validity and Validation*, in which she argues for the use of CFA as a tool to investigate the factor structure of assessment data, as well as to investigate construct-irrelevant variance and score dimensionality. Further, the Standards for Educational

and Psychological Testing list CFA as a source of evidence towards the validity of assessments (AERA, APA, NCME, 1999).

The APA affiliated journal *Psychological Assessment* was used to evaluate the percentage of articles within an assessment and measurement journal that use CFA as part of their research methodology. This journal was chosen because it is focused primarily on the development and evaluation of assessment/survey tools which would be a natural venue for publishing exploratory and confirmatory factor analyses. The journal's website states that topics relevant for publication consideration include: "development, validation, and application of assessment instruments, scales, observational methods, and interviews" (American Psychological Association, 2015). The previous four issues, constituting one year of publication (2014-2015), were examined to determine the prevalence of CFA publications. Out of 131 total articles in the one-year period, 43 articles (33%) used at least one CFA. For any given issue in that time period, the percentage of articles containing a CFA ranged from 26% to 47%. Clearly, in social science research focused on the evaluation of assessments/surveys, CFA is a commonly used methodology.

Examples in Education Research. There are many examples of the use of CFA in assessment, including some that account for method effects that cause error covariance. For example, Thurber, Shinn, and Smolkowski (2002) conducted a study of a multifactor model of reading and math with data from N = 207 fourth grade students, and explicitly took into account a method effect using a "Timed Test" method factor, since a handful of the indicators they used were timed measures. In the best fitting model, math applications (Factor 1) had three indicators, math computation (Factor 2) had seven, and reading (Factor 3) was

represented by three. The timed method factor loaded on eight indicators, five of which were computation variables and three of which were reading variables. In addition, the researchers specified two correlations between error terms: one within-factor correlation for two of the math computation variables, and one between-factor error correlation between applications and computations on the same test. The estimated correlations between the applications, computation and reading factors ranged from 0.76 to 0.83. Loadings ranged from 0.38 to 0.90 on applications and 0.54 to 0.91 on computations, with loadings lower for those items cross-loading on the method factor. The correlated errors were 0.17 for the within-factor relationship and 0.07 for the between-factor relationship. Significance levels were not provided for any estimates, however.

As another example of the use of CFA in assessment, Purpura and Lonigan (2013) evaluated the factor structure of different measures of informal numeracy skills in preschool age children. Twenty-five different assessments were used with *N* = 393 preschoolers to evaluate a theoretical 3-factor model. The authors compared the model fit of five separate models, consisting of a one-factor model, three two-factor models, varying combinations of factors, and a three-factor model. Model modifications were made based on changes in modification indices to improve model-to-data fit. Modification indices (MI) give the expected change in the overall model chi-square if a parameter currently constrained to zero is freed (Kline, 2010). The Lagrange Multiplier is one type of MI reported by some statistical packages (Lee & Bentler, 1980). The 3-factor model provided the best fit to the data. Between-factor correlations were quite high, ranging from 0.80 to 0.88, and there was a variety of factor loadings, ranging from 0.53 to 0.90, with the mean loading equal to 0.72. Two within-factor

correlated error terms were estimated, equaling 0.44 and 0.33, accounting for common measurement variance for skills measured with the same task.

Norwalk, DiPerna, and Lei (2014) conducted a CFA study examining the factor structure of the Early Arithmetic, Reading, and Learning Indicators (EARLI), a brief assessment of early numeracy and literacy skill. A sample of Head Start preschoolers, *N* = 289, were given 12 short assessments, with six assessing numeracy skills such as Counting Aloud and Pattern Recognition, and six assessing literacy skills such as Alphabet Recitation and Letter Sounds. Each assessment's raw score was based on the number of correct responses given within a certain time frame. Consequently, each variable could be viewed as representing an interval scale, similar to the assumptions in the present study. The authors evaluated four CFA models for model fit, a 1-factor model, 2-factor model, a 2-factor model with correlated error terms, and a bifactor model. The bifactor model replicated the structure of the 2-factor model with the addition of a third "general" factor indicated by all of the observed variables. The need for correlated error terms on the 2-factor model was evaluated using a correlation matrix of error terms and the Lagrange Multiplier. A discussion of their effect on model interpretation or the possible theoretical reasons for their inclusion was not included in the article. In the 2-factor model with correlated error terms the factor loadings ranged from 0.44 to 0.85, with a mean loading of 0.67. The model included 6 within-factor correlated error terms ranging in magnitude from 0.12 to 0.38, with an average of 0.20. The error correlations affected a total of eight items, with only two variables on the numeracy factor, and the rest on the literacy factor.

These are just three examples of a multitude of CFA studies in education and social science research. There is wide variety in the number of latent factors and observed indicators

used in models. In any discipline where assessments and surveys are used to evaluate or validate constructs that cannot be directly measured, the estimation of latent factors in CFA models is likely to be used.

Reflective Measurement. As already mentioned, latent variable models (aka structural equation models) such as CFAs assume that the observed measures/items are representational of the unobserved latent construct(s). In other words, the scores from observed measures/items can be viewed as indicators of the unobserved, latent construct(s). Kline (2012) terms this idea "reflective measurement," which makes explicit the assumption that the latent construct causes the performance on observed measures, not the reverse. He states "Most measurement models analyzed in SEM studies—and all models tested with CFA—assume reflective measurement"(p. 112). There are some instances where the latent variable is defined as a composite of the indicator variables, where the variables cause the construct, socio-economic status for instance, rather than some underlying trait giving rise to answers on the observed variables. However, the assumption of reflective measurement is far more common in measurement models, and was assumed throughout the current study. This said, the observed variables are not assumed to be perfect measures of the latent construct. An individual's score on a given variable is dependent not only on his or her true trait/knowledge/ability, but also on error present in the observed measures.

Measurement Error. In traditional "univariate and bivariate statistics, multiple regression and multi-equation systems," the observed variables are treated as perfect measures of their corresponding latent variable, failing to account for any error in the measurement (Bollen, 1989). Unfortunately, there exist many potential sources of error in any

representational measurement where the true construct is unobserved, such as math ability, feelings of support, and achievement motivation. Sources of measurement error could include sources as simple as a noisy testing environment or a student's physical well-being on the day of the assessment, as well as more complicated factors associated with test construction. Importantly, in his seminal text, *Structural Equations with Latent Variables*, Bollen (1989) showed empirically and experimentally that the failure to account for measurement error in models where latent variables are treated as observed variables can result in both positively and negatively biased parameter estimates, depending on the variable's relationship to the error variance. To illustrate this, Bollen used reliability estimates to estimate the error variances in a path analysis involving two observed variables and three latent variables, each estimated from a single indicator. He then estimated the model parameters using two different estimation methods while also varying whether or not the error variances were taken into account. The results make it clear that failing to account for measurement error in structural equation models biases parameter estimates.

Mathematically, measurement error is defined as the difference between a measured value of the construct and the true value of the construct. In classical test theory (CTT), a subject's measured score, X, is assumed to be a combination of his/her true score on a given trait, T, plus measurement error, E (Allen & Yen, 1979), giving the equation:

$$X = T + E \tag{1}$$

In this case, measurement error is assumed to be purely random. Similar to CTT, in any latent variable model, variability in each observed variable is assumed to be caused by both the latent variable, due to reflective measurement, as well as measurement error. In most cases,

measurement error is explicitly estimated as a surrogate variable for each manifest observed variable. In other words, “Measurement errors are proxy variables for all sources of residual variation that are not explained by the model” (Kline, 2010, p. 113). Due to this, measurement error in SEM is defined as a latent variable (Bollen, 2002). The error is not observed, like the score of the observed variable, but rather it is estimated in concert with the latent factor.

While CTT assumes that the sources of this error are random, there is no such assumption in SEM. In fact, the measurement error term in SEM represents two types of unique variance: “*random measurement error* (in the psychometric sense) and *error uniqueness*, a term used to describe error variance arising from some characteristic considered specific (or unique) to a particular indicator variable”(Byrne, 2006, p. 10). Thus, other factors that influence a subject’s score on an observed variable, besides the modeled latent factor of interest, are contained within the error term for that observed variable. Byrne’s statement also hints at the nature of this non-random error contained within the term.

The Assumption of Linear Independence

In CTT, the errors for each item are assumed to be independent of, and thus uncorrelated with, each other. Many aspects of CTT, coefficient alpha for instance, are based on the assumption of uncorrelated errors among items (cf., Green & Hershberger, 2000). The assumption is also important for most SEM models, and relatedly, for item response theory models as well (Hambleton, Swaminathan, & Rogers, 1991).

As previously stated, the measurement error in SEM is composed of both random error and variability caused by other factors not contained within the latent construct. Generally, we assume any unmodeled factors affecting one observed variable are independent from the

unmodeled factors affecting other observed variables in the model. Observed variables can be single items on a test or survey instrument, or total scores from many assessments. In either case, we assume that, once the main factor(s) influencing responses to the observed variables are partialled out, those responses then become uncorrelated, no longer sharing any variation. This assumption of the independence of errors is called "local," "linear," or "conditional" independence (Kline, 2012, p. 115). To reiterate, any unexplained variation in an observed variable should not correlate with another observed variable's unexplained variance (Heene, Hilbert, Freudenthaler, & Bühner, 2012).

This assumption has not always received the attention it merits. Even so, Heene et al. (2012) wrote, "It is crucial to recognize that the statistical definition of latent variables in the context of factor analysis is based on an assumption of linear independence once the latent variables (i.e., the common factors) have been specified" (p. 37). In a standard measurement model, as defined by Kline (2010), any covariation among indicators is left out of the model precisely because all unspecified causes of the observed variables are assumed to be independent and are thus left unanalyzed by the model. A standard measurement model is illustrated in Figure 1a. Specifically, the model assumes two latent (unobserved) factors indicated by four measured (observed) variables each, with errors assumed to be independent of each other. Figure 1b, in contrast, illustrates the same model but with a correlation path estimated between two of the first factor's indicators' error terms, accounting for a violation of linear independence among the two indictors.

According to Kline, the assumption of linear independence is "both restrictive and probably unrealistic" given the nature of measured outcome variables in the social sciences

(2012, p. 115). In noting this, Kline also argued that it is best to specify correlations among the error terms when there are substantive reasons for the associations. The specific effect of under-specified error correlations, when correlation among errors is present but not modeled, is discussed next.

Violations of the Assumption of Linear Independence

As previously discussed, violations of linear independence occur when indicator variables share variance due to some factor other than the specified latent factor of interest. This unwanted component is sometimes referred to as construct-irrelevant variance; that is, variation in the observed variable that is unrelated to the latent construct of interest. If this error component, such as item format or wording, is shared among observed variables, then the variables are said to share error covariance. While the presence of error covariance among specific pairs or sets of observed variables may not affect the estimated score on the latent variable, the covariance might affect the underlying measurement model. There is some indication in the prior research that the failure to specify correlated error terms in SEMs that have error covariance can indeed result in biased model estimates and lead to an incorrect evaluation of the connection between the observed variables and the latent variable (Cole, Ciesla, & Steiger, 2007). This research is discussed in further detail later. The effect of error covariance can also be seen in estimations of reliability, such as coefficient alpha, spuriously inflating the estimates (See Cronbach, 1951; Fleishman & Benson, 1987; Miller, 1995; Rozeboom, 1989; Wolf, Harrington, Clark, & Miller, 2013; Zimmerman, Zumbo, & Lalonde, 1993). However, the effect on the estimates of loadings between observed variables and latent constructs as well as the between-factor correlations is the focus of this paper.

Potential Causes of Violations to the Linear Independence Assumption

As stated earlier, observed variables involving representational measurement are measured with error. The error in a particular observed variable can be divided into two sources, purely random error, and the possibility of "some unwanted component that is stable across successive measures"(Gerbing & Anderson, 1984, p. 573). This latter type of error, called construct-irrelevant variance, can be due to unintended factors occurring within a given survey or assessment instrument, including systematically different item response formats, such as multiple-choice, true/false, and free recall responses, as well as systematically different item stem wording, such as negatively and positively worded stems. In addition, item order effects can also cause a construct-irrelevant variance. In sum, observed variables with systematic response patterns unrelated to the construct are said to share error covariance, and correlated error terms among affected variables should be specified. While the presence of unmodeled error covariance among specific pairs or groups of variables may not affect the observed total scores for individual subjects, they may affect estimates of the underlying model, including the relationships between observed variables and the latent factors and/or between latent factors.

Briefly, the measurement part of an SEM model is used to assess the relationship between the underlying theoretical factors and the observed variables used as indicators, and the structural part of an SEM is used to assess the relationship among the theoretical factors themselves. Additionally, the fit of the data to the specified model can be assessed, indicating whether the tested model explains the variable-factor relationships well. Research has shown that indicators of model fit are not always sensitive to misspecification of correlations among error variances (Heene et al., 2012). That is, the relationships between the error variances of

observed variables that are not accounted for in the model may not cause a poor model fit. However, this type of misspecification can have an effect on model loading estimates and subsequent interpretation (Cole et al., 2007). Consequently, a review of possible causes of error covariance is important, as they could affect the underlying model for the assessment.

Method effects. The majority of methodological research on the causes of construct-irrelevant variance on observed variables appears to have been conducted in relation to surveys (see Podsakoff, MacKenzie, Lee, & Podsakoff, 2003 for a review). This research has indicated that order, language, reversed items, and other method effects can have an effect on how subjects answer survey questions (Schuman & Presser, 1996; Todorov, 2000; Tourangeau, Singer, & Presser, 2003; Uhan & Fink, 2013). In general, method effects are systematic differences in item responses that are due to item formatting or placement in a survey or assessment instrument. In Marsh and Bailey's words, method variance is defined as "an undesirable source of systematic variance that distorts correlations between different traits measured with the same method" (1991, p. 48). Maul (2013) noted that the differences in subject responses to items due to survey modes, including paper-and-pencil, online-based, and in-person interviews, can also constitute a method effect. In all of these cases, the subject's responses to the particular observed variable may depend on construct-irrelevant factors.

As one example, Harvey, Billings, and Nilan (1985) evaluated the factor structure of the Job Diagnostic Survey using a CFA model. Their research showed that the inclusion of a negative wording factor, for survey items where a low score indicated a high amount of the latent construct, significantly improved the fit of the model. This result led the researchers to posit

that the format of the negatively worded survey items represented a separate latent factor in addition the main latent factors of interest.

Another example of method effects in survey research involves what are called “context effects.” A context effect is defined as “a process in which prior questions affect responses to later questions in surveys” (Holyk, 2008); in cognitive psychology, this effect is also known as a priming effect whereby the presence of a stimulus word evokes associations with other words that are closely associated perceptually (e.g., Thompson-Schill, Kurtz, & Gabrieli, 1998). As an early example of this effect, Bishop, Oldendick, and Tuchfarber (1982) conducted a field experiment to investigate how the different location and subsequent context of a survey item might change responses. An item on political interested was varied in location, either immediately after a set of items concerning the upcoming election, or immediately before. Respondents were also given a question to evaluate their knowledge of the upcoming election. The results showed that reported political interest did vary according to the location of the survey item, however there was an interaction effect with knowledge of the upcoming election. The more respondents knew about the election, the less likely the location of the item would affect their response.

Clearly, assessments can easily be affected by method effects as well. For example, in a test for pharmacists, Caldwell and Pate (2013) found that non-standard item formats were significantly more difficult for students, while providing no better discrimination; discrimination being defined as the precision with which an item can differentiate between students with different abilities. Many of the non-standard items under investigation in the study, including negatively worded stems, number of response option distractors, and the use of the “none of

the above" response option, are commonly acknowledged as item formats to avoid or use with caution (Dillman, Smyth, & Christian, 2009; Taylor & Nolen, 2005). Even so, different response patterns due to these method effects is likely indicative of some relationship among items that is not accounted for by the latent construct, thereby inducing error covariance among those items.

Math assessments are particularly susceptible to a variety of method effects due to the nature of multi-step processes involving complex response options. Complex response options can be thought of as response options that enable students to present answers in a variety of manners, such as graphs, tables, equations and sentences/paragraphs, and involve a more in-depth representation of the answer. Graf (2009) presents "complex response types" as a way to more deeply delve into student understanding. "Complex responses can provide valuable evidence about students' process competencies" (p. 52). However, at the same time, if a student does not use the expected method for solving the problem/item, then he/she may embark on a more drawn out process, resulting in a loss of time for completing the other questions. It is possible to then modify the problem/task to provide guidance to the student in selecting a problem-solving method, however, this may detract from the information about the students' competencies, defeating the purpose of a complex response format. Relevant to the current discussion is that there is also the possibility that the response format will cause covariance among items unrelated to the latent construct.

In addition to problem-solving and response method effects, consistent vs. inconsistent language is sometimes used in the construction of mathematics test items. Some items may employ language that is suggestive of the types of operations necessary to solve the problems,

or what is known as consistent language. In the same test, however, other items may be worded with language that is consistent with one representation of the problem, while the actual problem is consistent with a different representation. This mismatch is referred to as inconsistent language (Lewis & Mayer, 1987). Items with inconsistent language may be more difficult than a problem addressing the same concept with consistent language. The increased difficulty among items with inconsistent language may then cause covariation beyond that of the latent variable, constituting a violation of linear independence. Hence, "inconsistency of language" would require specifying correlated error terms in a CFA model, as this factor would not be accounted for by the latent mathematics knowledge construct.

Method effects on measurement model estimation have also been extensively researched in the multi-trait/multi-method (MTMM) line of research (e.g., Kenny & Kashy, 1992; Marsh & Hocevar, 1988). For example, Marsh and Bailey (1991, p. 48) defined one method effect in their data as the four waves of data collection used to gather subjects' responses.

Clearly, there is high potential for method effects to occur in items from survey and assessment instruments. Irrespective of whether the effects are unintended or planned, applied researchers may not be aware that they are violating the assumption of linear independence when these effects are ignored in their models.

Other sources of construct-irrelevant variance. Other research also implies the possible presence of item relationships or correlations based on factors other than the primary construct or latent factor. Taylor and Lee (2011) conducted differential item functioning (DIF) analysis on ethnic groups for a large scale criterion-referenced state achievement test for

reading. DIF analysis seeks to evaluate differences in item parameters, difficulty or discrimination, based on a student's inclusion in an ethnic group compared to another focal ethnic group. The authors found that items were flagged for DIF were those that had specific response formats. For example, items with multiple-choice formats generally favored White students, whereas items with constructed-response formats favored Asian/Pacific Islander, Black/African American, and Latino/Hispanic students. In a separate study of DIF on math test items due to gender, Taylor and Lee (2012) found similar results: items with multiple-choice response formats favored male students, whereas items with constructed-response option formats favored female students.

The incidence of error covariance due to a shared item format can also be design-driven, resulting from the nature of the study. CFAs on longitudinal data commonly contain correlated errors across time points (Little, 2013). When the same instrument is used to measure the same latent factor at different time points, the error associated with each indicator variable may be stable. While linear independence may be tenable for all items within a given instrument used to measure subjects at a single time point, linear dependence most certainly will not be tenable for measures used repeatedly over time. Due to the prevalence of situations potentially inducing error covariance, further research on the effects of violations of the assumption of linear dependence is needed.

Prior Research on SEM Model Misspecification

There have been several studies evaluating the impact of model misspecification in SEMs (e.g., Fan & Sivo, 2007; Heene et al., 2012; Hu & Bentler, 1998; Saris, Satorra, & Van der Veld, 2009). Model misspecification can result from either over- or underparameterization.

Overparameterization occurs when a path whose population value is zero is specified in the model, adding an additional, unnecessary parameter estimate to the model. Underparameterization can be considered the opposite of overparameterization. In underparameterization, paths whose population values are non-zero are instead specified as zero. This misspecification leaves out the estimation of population parameters that should be part of the model. In many instances, this type of misspecification, underparameterization, is of main interest to the researcher.

Much of the work on model misspecification to date has centered on the sensitivity of model fit indices (FI) to misspecifications. This is a fertile field of research given all of the possibilities concerning types of models, types of misspecifications, and magnitude of misspecification, to name a few. Given that the SEM framework is very flexible for evaluating innumerable different models, the types and levels of misspecifications available for study are numerous, especially given research showing that all of these factors potentially have a different effect on model FI (Fan, Wang, & Thompson, 1999; Hu & Bentler, 1998; Saris et al., 2009).

For example, Hu and Bentler (1998) conducted a simulation study looking at the sensitivity of fit indices to underparameterized model misspecification. In a three factor confirmatory factor model with five indicators per factor, four separate misspecified models were tested. Two “simple” models were specified with misspecifications to the factor covariances, and two “complex” models were specified with non-zero path loadings, cross-loading on an additional factor, specified as zero. For each model, seven distributions, six sample sizes, and three estimation methods were evaluated. The estimation methods were

maximum likelihood (ML), generalized least squares (GLS) and asymptotic distribution-free method (ADF). The authors studied a total of 15 fit indices (FI), including the normative fit index (NFI), Tucker-Lewis index (TLI), comparative fit index (CFI), standardized root-mean-square residual (SRMR), and root-mean-square error of approximation (RMSEA). Overall, fit Indices based on ML estimation performed better than the other estimation methods. Within the ML estimation, clusters of FI were found to be highly correlated. NFI was clustered with a group of indices, while TLI, CFI and RMSEA were clustered in another group. SRMR was found to be the least similar to other FI. The authors also found that SRMR was most sensitive to "simple" misspecifications, while TLI and RMSEA, among others, were more sensitive to complex misspecifications. Sample size had a small effect on NFI, TLI, CFI and SRMR, but a large impact on RMSEA. It is worth noting that the factor loadings and correlations among latent factors used in this study were relatively large, with loadings ranging from .70 to .80, and correlations equal to .30, .40, and .50. The importance of these magnitudes of parameter estimates are shown in subsequent research, which found interactions between sample size and loading magnitudes on fit indices (Heene et al., 2012).

In an extension of Hu and Bentler's (1998) study, Fan and Sivo (2005) simulated multivariate normal data and used two estimation methods to evaluate the effect of model misspecification on FI. The authors used two models, two conditions of misspecification (slight and moderate), and under each condition, two levels of misspecification. The authors also varied the sample sizes ranging from 150 to 1,500, at intervals of 150. The two models were the same as the Hu and Bentler's original study: one simple and one complex. The purpose of the new study, however, was to control the severity of model misspecification, as the previous

study's types of misspecification and severity of misspecification were confounded. The severity of the misspecification was defined by the change in the noncentrality parameter, the noncentral χ^2, due to the misspecification. Fan and Sivo concluded from their study's results that the 2-index strategy proposed by Hu and Bentler was not supported. They found that the SRMR was not more sensitive to factor covariance misspecifications than other types of FIs. Thus, the sensitivity of FIs to various types and magnitudes of SEM misspecifications remains a question open to study.

In a subsequent study, Fan and Sivo (2007) re-evaluated the effects of model misspecification on the sensitivity of FI, but also incorporated variations in model types. Three models were tested: two CFA models, mirroring the same CFA models used by Hu and Bentler (1998) and one SEM model with both exogenous and endogenous variables. Exogenous variables can be thought of as the independent variables, they're causes are not incorporated in the model; the latent factors in CFA analysis are exogenous. Endogenous variables have their causes explicitly represented in the model, similar to the familiar dependent variable. Three levels of specification error were models, with two levels specification error severity. The misspecification severity was quantified by estimating the power to reject the misspecified model under the given sample size. The results of their study indicated that model FIs are sensitive to model type, but not misspecifications. This result was supports a previous study by (Fan et al., 1999) that found wide variation in the performance of FI under various conditions of sample size, estimation method and model specification.

Even though much of the research in model misspecification in SEM has focused on the effect on model fit, there is awareness that the misspecification can lead to other problems

throughout the model. Cole (1987) noted that model estimates can be fairly inaccurate if the specified model does not match the underlying model in the data, even though this type of misspecification can still lead to good fit of the model to the data. Additionally, there are usually multiple equivalent models equal in model fit to the correct specification matching the underlying data, a fact often ignored in publications (MacCallum, Wegener, Uchino, & Fabrigar, 1993). Consequently, model misspecification can be seen as an important area for continued research in SEM.

Correlated Error Terms and Specification Searches

In most of the studies conducted on SEM misspecifications, researchers have focused on the misspecification of either factor loadings or structural paths, like between-factor correlations or path loadings between latent factors. Factor loading and structural path specifications are based on research and theory for the particular model, and so too can the specification of correlated error terms. However, this is not always the case. Researchers may be unaware of the conditions potentially violating the assumption of linear independence, or they may be hesitant to specify correlated errors.

Throughout the literature, a multitude of warnings can be found cautioning the researcher from the post-hoc addition of correlated errors to the model, the addition of error correlations after the model has previously been specified and analyzed (Hoyle & Smith, 1994). The problem with the use of specification searches “is that researchers may be capitalizing on random, sample specific characteristics of the data” (Cole et al., 2007, p. 381). This echoes the sentiments of Gerbing and Anderson (1984) who also questioned the effect of post-hoc addition of correlated errors and their effect on the substantive interpretation of the model.

Adding correlated errors to the model for the sole purpose of improving model fit without any substantive rationale not only detracts from the meaningful interpretation of the model, but also runs the risk of being sample specific, failing to generalize to the population.

Evidence of sample specificity problems comes from MacCallum, Roznowski, and Necowitz (1992), who looked at the effects of post-hoc, iterative specification changes on model fit, and their generalizability to other samples from the same population. The authors used an initial model and data set to conduct a sampling study where ten pairs of random samples for each level of sample size were pulled from the original data set. Samples of size 100, 150, 200, 250, 325, 400, 800, and 1,200 were used for the study. Pairs of samples were pulled from the data so that one could be used as a calibration sample, while the other was used as a cross-validation sample. Once the model was fit to the calibration sample, a number of sequential modifications were made to the model using modification indices. The final model obtained through modifications from the calibration sample was then checked against the cross-validation sample. The results of the study showed that even when sample sizes are large, model modification searches may not be stable across samples. In fact, in small to moderately large samples, there is substantial evidence that the sequential modification search method is highly unstable, resulting in little generalizability to other samples in the population, or to the population as a whole. These results should give researchers some pause when specifying correlated error terms among observed variables.

All this said, correlated error terms need not be added based on post-hoc specification searches. Instead, there are often logical or substantive reasons, based on prior research and

knowledge of variable measurement formats, to specify error covariances among certain observed variables in the model.

Alternative Models for Error Covariance

The specification of correlated error terms need not be the only method for accounting for error covariance in CFA models. Gerbing and Anderson (1984) investigated the use of correlated errors in CFA models, and presented an alternative method of model specification to account for many of the same aspects as correlated errors. The authors argued against the unchecked use (i.e., post-hoc searches) of correlated errors in CFA models for more than just the possibility of capitalizing on chance to improve model fit. The authors argued that simply adding a correlated error term to improve model fit to an acceptable level, while failing to further explore possible adjustments to the model specification, could lead to an improper model. They argue that adding a single correlated error could mask the fact that the data is better represented by another model. Instead, they posited that operationalizing the factor of interest as a second-order factor, with the first-order factor(s) arranged in unidimensional groupings, may better represent the data. The first-order factors can be further decomposed into the influence of the second order factor and a component unique to that first-order factor. "From the perspective of estimating the second-order factor, the group specific component is irrelevant, stable error—an unwanted component of a given first order factor"(p. 576). Thus, the second-order factor structure is able to account for construct irrelevant effects, while perhaps providing a more meaningful structural alternative to correlated errors among indicators within factors.

The structure of correlated errors within CFA models is not always present as within-factor covariation. Method effects and other sources of correlated error terms can occur between factors. Consequently, a second-order CFA model is not always an appropriate method to account for covariation among error terms. At such times, it may be more proper to specify a correlation among the error terms of two observed variables across factors that are thought to share some common method effect.

The Effect of Under-Parameterization of Correlated Error Terms

While the research concerning the sensitivity of model fit indices to structural and measurement model misspecifications was reviewed earlier, research on the effects of the misspecification of correlated error terms has seen little work. The ability to use correlated error terms in SEM has led to no shortage of arguments in the literature. Cole et al. (2007) went so far as to write "The capacity to allow and test correlations between residual terms is simultaneously one of the greatest strengths and most worrisome dangers associated with latent-variable covariance structure analysis"(p. 381). In their study, Cole and colleagues specified the population covariances for a set of date, then they were able to calculate results for the just-identified models. They specified a baseline model of six indicators from two methods loading on a single latent factor. The errors for the three indicators within each method all correlated within the model, representing a method effect, similar to the correlated trait/correlated uniqueness (CTCU) model in multi-trait multi-method research (MTMM) (Marsh, 1990; Marsh, Byrne, & Craven, 1992). When models were estimated with two indicators removed from one method and one from the other and the correlated errors unspecified because the subsequent model would be under-identified, the resultant parameter

estimates were biased. As seen in Figure 2, adapted from Cole et al, 2007, the bracketed population estimates differ substantially from the model estimates, which are not bracketed. Observed variables sharing error variance were subject to positive bias in factor loadings due to the under-specification of the correlated error term, while the item unaffected by the error covariance showed negative bias. The change in loadings then lead to erroneous interpretations of the strength of the relationship between the observed variables and the latent factor. Thus, failing to account for correlated errors due to research design and the number of indicators collected or problems with model identification can bias the estimates and potentially cause faulty conclusions.

The Cole et al. (2007) study examined design-driven correlated error terms. Specifically, the correlated error terms were part of the research design, based on a shared method effect between observed variable groupings. In situations where there are strong theoretical reasons to include correlated errors in the measurement model, it appears that the analysis should allow for their specification, with the constraint of maintaining model identification. Sometimes the theoretical reasons for including correlated error terms in the model specifications may not be clear to the researcher or even clear based on the given research. It is generally accepted that the over- or under-specification of structural or measurement model components will take away from the substantive conclusions drawn from the model. Whether the given model is meant to be an explanatory model, or look at the factor structure of a given measure, the failure to include necessary paths or the inclusion of unnecessary paths can significantly alter the interpretation of the model.

While the Cole et al. (2007) study provides one case for the dangers of leaving out correlated errors when the design calls for their use, a consensus on the effect of failing to include correlated error terms has still not been entirely reached. While numerous authors strongly caution against the post-hoc search for and use of addition of correlated errors, little research other than the Cole et al (2007) study has been conducted to shed light on the effect of not specifying correlated errors when they are present.

Finally, the most recent research study on the effect of misspecified error terms should be reviewed. Specifically, Heene et al. (2012) generated data for a 2-factor CFA, with 12 indicators for each factor, to test for sensitivity of fit indices (FIs) to misspecification of correlated errors among indicator variables. The authors varied the number of correlated errors present in the simulated data between three and six, representing a mix between positive and negative, -0.3 and 0.3, and all positive, 0.3 and 0.5. The correlated error terms were between factors, allowing the error terms for indicators on different factors to covary. The misspecified models were those specified with no correlated error terms; models failing to account for the correlations among errors. The sample sizes used for analysis and the magnitudes of the factor loadings were also varied in the study. Two levels of factor loadings were used. In the first loadings ranged from .31 to .59, while the second level added a constant .20 to all of the loadings. The sample sizes were set at N = 150, 250, 500, 1,000 and 2,500. The analysis showed that, under almost all sample sizes and levels of misspecification for the smaller loadings, the RMSEA and SRMR values stayed near or below the generally accepted cut-off values. In contrast to this, The CFI nearly always indicated poor fit, except at large sample sizes ($N \geq 500$), or small misspecifications (three positive and negative). For the larger loadings, the FIs were shown to

be even less sensitive to misspecification. They did not examine loading bias or other parameter estimates, however; as such, it is unclear how the misspecifications affected those values.

The results of the Heene et al. (2012) study provide additional evidence that FIs are not always sensitive to the types of error term misspecifications. Further, the research has shown that specification searches can lead to sample-specific results (MacCallum et al., 1992), and he post-hoc additions of correlated errors for the purpose of improving model fit can damage results interpretation (Cole et al., 2007; Hoyle, 1995). On the other hand, one prior research study has also shown that the misspecification of theoretically based correlated errors due to research design can bias parameter results (Cole et al., 2007). In sum, it would appear that careful consideration of the data and the model are especially important and that additional work in the area of parameter bias is sorely needed to help researchers understand the consequence of not using correlated errors when it may be warranted.

Present Study Research Questions

Structural equation modeling (SEM), and in particular, confirmatory factor analysis (CFA), play an important role in research, particularly in the field of educational measurement and evaluation. Though the method for estimating the relationship between observed variables and latent variables has been an important tool for many years, the complexity of the models and the related assumptions has left many potential influences on the model under researched. Given the multitude of possibilities for the creation of error covariance among observed variables in surveys and assessments, further evaluation of the potential effect of the misspecification of the correlated error terms necessary to account for the error covariance is needed. This leads to two major research questions:

1) How does the failure to account for error covariance among items with correlated error terms in the model effect parameter estimates in confirmatory factor models?
2) How does the presence or absence of correlated error terms substantively change model estimates for data in which significant error covariation is hypothesized?

Chapter II: Monte Carlo Simulation

Monte Carlo (MC) studies are statistical sampling studies where multiple data sets are generated and analyzed to study the impact of some condition on the estimates in a model. The conditions commonly studied in MC simulations are enumerated later in the paper. In structural equation modeling (SEM), the flexibility of the framework and the complexity of models make it a fertile ground for MC research. Another reason for the wealth of simulation studies in SEM is the lack of analytical proofs for many of the problems. If an analytic proof can provide the answer to a problem, then a MC simulation is not necessary; as a proof provides the answer across all possible conditions. However, analytic proofs are only possible if the mathematics of the situation have been worked out. There are many statistical situations where the mathematics has not be worked out and may never be (Mooney, 1997). Frequently, the complexity of the problem prohibits a possible analytic solution. Robert and Casella (2013) note that latent variable modeling, i.e. structural equation modeling, is a good example of when this is true.

An important use of MC simulations is to investigate conditions that affect applied research. Paxton, Curran, Bollen, Kirby, and Chen (2001) noted that the lack of external validity is one of the major reasons that MC simulations are criticized. Because of this, an MC simulation should not be undertaken without a "compelling rational for using these techniques"(Harwell, Stone, Hsu, & Kirisci, 1996, p. 103). When MC research is used to investigate problems that affect applied research, the conditions of the simulation should mimic the conditions of the applied research as closely as possible. Otherwise, the results of the MC study are of little practical use to applied researchers (Bandalos & Leite, 2013). If the basic

conditions set forth in the study bear little resemblance to conditions found by applied researchers, then the results of the research are unlikely to provide any guidance to those same applied researchers. Thus, it is important to closely ground MC simulations in applied situations.

General Application of MC Simulations

Monte Carlo simulation is a widely employed technique for studying the effect of a defined research condition on particular statistical outcomes of interest. Paxton et al. (2001) defined nine steps in planning and executing a MC study as follows (p. 288):

1. Developing a theoretically derived research question of interest
2. Create a valid model
3. Designing specific experimental conditions
4. Choosing values of population parameters
5. Choosing an appropriate software package
6. Executing the simulations
7. File storage
8. Troubleshooting and verification
9. Summarizing the results.

Having previously defined the research question, each of the subsequent steps defined by Paxton et al. is followed in the Methods and Results sections of this paper. However, further elaboration of the steps to conducting a MC study are worth discussion.

The decision of what model to evaluate in the present study is based on prior research's models (namely, Cole et al., 2007) but the basic model structure has relevant applications to current educational measurement research, particularly with regard to the development of new assessment systems for the Common Core State Standards for Mathematics (CCSSM). One of these systems of assessments is the Smarter Balanced Assessment System (SBAS). While the actual model used for the SBAS structure is unnecessarily complex for the current research study, it serves a backdrop for connecting the simulation results to that of real-world context.

Once the model to be evaluated is determined, then the parameters for generating the "pseudo-population" must be decided upon. "This pseudo-population consists of mathematical procedures for generating sets of numbers that resemble samples of data drawn from the true population" (Mooney, 1997, p. 4). The parameter values could be based on a review of the applied literature to determine the range and variability of the values in applied research. The data for the MC study are then generated based on values representative of the results generally found in applied studies. Another approach, recommended by MacCallum (2003), involves treating a single extant data source as the pseudo-population. The estimates for the extant data source are treated as population values, and then the data are generated from these population values. Thus, the simulation more closely mirrors the characteristics of applied research. MacCallum argued that studies based on this second approach are of greater use to practitioners. This said, when evaluating a complex situation for which little research exists (as is the case in the present study), it may be best to restrict the parameter values to a plausible range based on prior research both in applied settings and MC simulations. While loosely matching applied research estimates, the chosen pseudo-population values for the current study can still provide insight to applied researchers. At the same time, more closely matching population values from previous MC simulation research helps to replicate and extend those prior research findings. Thus, the range of sample size, loading and between-factor correlations were meant to replicate previous MC research while still maintaining relevance to applied research. The model chosen for the current study was partially based on the model used in Cole et al. (2007), but also provided a simple, yet plausible model to evaluate the potential effects of unmodeled correlated errors.

Application of MC Simulations in SEM

Monte Carlo simulation has been used to evaluate a wide variety of situations within structural equation modeling. Bandalos and Leite (2013) examined the application of MC studies in SEM. They noted many of the situations where MC simulation is useful, including:

- Violations of assumptions such as:
  - Normality
  - Independence of observations
  - Model misspecification
  - Linearity of effects
- Investigation of issues such as:
  - Small sample behavior of asymptotically-based statistics
  - Continuousness of observed variable distributions
  - Effects of estimating large numbers of model parameters
  - Effects of analyzing correlation matrices rather than covariance matrices
  - Properties of goodness-of-fit statistics for which no mathematical distribution exists (pp. 627-628)

Even when many authors have produced MC studies concerning evaluations of different aspects of SEMs, the overall complexity of SEM often leaves questions unanswered, conditions untested and models unstudied. “For example, historically the bulk of the MC studies in SEM have been conducted with CFA models; in a review of 62 Monte Carlo studies in SEM, Hoogland and Boomsma (1998) found that 89% used CFA Models”(Bandalos & Leite, 2013, p. 632). Despite the prevalence of CFA models within SEM MC research, there are still conditions left under-researched.

Chapter III: Monte Carlo Simulation Study Methods

The current study aimed to evaluate the impact of an omitted correlated error term in a 2-factor confirmatory factor analysis model, a popular latent variable model in measurement and evaluation. A simple 2-factor CFA, with four unit normal indicators loading on each factor, was used to facilitate understanding the conditions under which a single within-factor correlated error may impact model validity. Figure 3 illustrates the general model employed for the present study.

General Approach

A Monte Carlo (MC) simulation study was conducted using Mplus 7.2 (Muthén & Muthén, 1998-2012). In this software, data is generated using a correctly specified population model, which is then analyzed for parameter estimation using a misspecified model. For this study, maximum likelihood estimation, the default in Mplus, is used. The correctly specified model included correlated error terms, while the misspecified model omitted the correlated error term. This cycle repeats until the desired number of replicates are simulated and analyzed. In the present study, 100,000 replicates were used for each model condition. As computing power and accessibility has increased, so has the ability to conduct Monte Carlo simulations. The large number of replicates in this study was made possible by the availability of the necessary computing power and allows for the increased stability of the mean model estimate for each condition.

The MONTECARLO command in Mplus does not include the ability to conduct loops across multiple conditions. Loops allow for all of the conditions to be defined, crossed and analyzed sequentially. Appendix I displays example code for one condition. Due to the lack of

this ability in *Mplus*, the MplusAutomation package (Hallquist & Wiley, 2014) in R (R Core Team, 2013) was used to replicate the ability to run loops. The package created multiple input files crossing all of the model conditions. Each input file was then automatically executed inside Mplus, using a single command in *R*. Appendix II displays example code for this package. Finally, R was used to read the estimated parameters from each output file, and aggregate and organize them into a single data file. The data file was further organized using the R package "Reshape2" (Wickham, 2007). Example code is given in Appendix III. The data file was then exported to SPSS 19 (IBM Corp., 2010) for secondary analysis to test for condition main effects and 2-way interactions on bias in loading estimates as well as 95% confidence interval coverage.

Experimental Conditions

The model analyzed consisted of several variations of a 2-factor CFA model, with four unit normal indicator items/variables per factor. This type of model, sometimes referred to as the standard CFA model, is cited by Kline (2010) as one of the most commonly tested CFA models in the literature. This model, being relatively standard, provided clear opportunities to assess the relationship between any bias that may present and the model characteristics, with limited opportunities for confounds. Simply put, the simplicity of the model allows for unfettered analysis.

Since no prior research has systematically examined the impact of correlated error on model parameter estimates, for the present study correlated error was induced between two items within the first factor, otherwise known as a within-factor correlated error term. Additionally, although item-factor loading magnitudes were varied across conditions, within

each replicate, all indicators were restricted to have the same loading value for purposes of simplicity. The fully crossed conditions that were varied for this study included sample size, magnitude of the between-factor correlation, magnitudes of the item-factor loadings, and magnitude of the within-factor error correlation, yielding 360 total condition combinations. Specific manipulated conditions were as follows.

Sample size. Three levels of sample size were used, small (N = 80), medium (N = 160) and large (N = 400). The sample sizes selected for study were based on N:p (sample size : indicator variable) ratios of 10:1, 20:1 and 50:1, respectively. These sample sizes are common in other Monte Carlo SEM research (Fan et al., 1999; Hu & Bentler, 1999) and can also be readily found in recent substantive applied research (Norwalk et al., 2014; Purpura & Lonigan, 2013; Richards, Levesque-Bristol, & Templin, 2014), with the exception of the smallest condition.

The body of research on recommended sample sizes for SEM estimation varies tremendously, depending on conditions within the model as well as whether model fit or model coefficient tests are of focal interest. Bentler and Yuan (1999) found that sample sizes as low as 60 were adequate for estimation of simple CFA models. For EFA models, which estimate more paths than CFA models, sample size : indicator variable ratios of 20:1 have been recommended for tests of loading coefficients (Stevens, 2009, p. 333), for which a 2-factor model with four indicators per factor would yield a total sample size of N = 160. In contrast, according to a power analysis for model fit using RMSEA, the necessary sample size to achieve an RMSEA ≤ 0.05 with 0.80 power and 19 degrees of freedom (for a 2-factor CFA) and $\alpha = 0.05$, when the actual model deviation is presumed to be RMSEA = 0.00, 0.01, and 0.02, are 455, 489, and 625, respectively (Preacher & Coffman, 2006, May). The levels of estimated model deviation from

the population values were included to test the range of possibilities in an applied setting. In practice, it is considered optimistic to assume zero deviance between the model implied and sample covariance matrices (Hancock & French, 2006). Thus, deviations in fit are included to inform the power analysis. Most recently, Wolf et al. (2013) investigated sample sizes across a number of different models. For 1-, 2- or 3-factor CFA models with three to eight indicators, there was a very wide range of sample sizes required to achieve sufficient power for testing loading coefficients. For example, the 2-factor models with four indicators per factor required $N > 450$ for detecting loadings of ±0.50, $N = 200$ for loadings of ±0.65, and $N > 100$ for loadings of ±0.80.

Clearly, the sample sizes used in the current study are all smaller than the sample sizes recommended by Preacher and Coffman (2006) for achieving low residual error to translate into good model fit, and the sample sizes recommended by Wolf et al. (2013) for detecting loadings with low magnitudes.

However, the present study's focus is on bias in parameter estimates (rather than hypothesis testing of loadings or achieving good model fit), and as such, employed sample sizes on the lower end of plausible sizes to specifically test the impact of correlated error on models with the bare minimum number of subjects suggested by prior MC and applied research.

Item-factor loading. Five factor loading magnitudes were used, keeping all factor loadings within a given model equal, at 0.40, 0.50, 0.60, 0.70, and 0.80, to help evaluate whether the error misspecification might have a joint effect with loading magnitude on model parameter estimates. The chosen loading magnitudes are representative of the range of loadings commonly found in applied research, especially studies concerning the validation of

factor structure within tests and surveys (e.g., Cole, 1987; Matos, Pinto-Gouveia, Gilbert, Duarte, & Figueiredo, 2015; Norwalk et al., 2014; Purpura & Lonigan, 2013). In CFA models, it is necessary to set a scale for the latent factor which each of the model estimates are scaled in relation to. One of the estimates must be fixed, so that the other estimates' values are set in relation to the fixed estimate (Little, 2013). In this case, factor variances were set equal to one in the model, allowing each of the factor loadings to be estimated freely. Additionally, the error variances were constrained based on the magnitude of the factor loading. The error variances were set equal to 1 minus the square of the loading; for instance, the error variance for a loading of 0.40 would equal 0.84. Constraining the factor and error variances in this manner standardizes each of the factor loadings. Thus, each loading level in the present study represents a standardized loading.

Between-factor correlation. Four levels of magnitudes of correlation between the two latent variable factors were used, based on factor correlation sizes commonly referred to as 'small', 'medium' and 'large' (Cohen, 1988), as well as the null condition of no correlation: 0.00, 0.10, 0.30, and 0.50. This may also be interpreted as the factors sharing 0%, 1%, 9%, and 25% variance, respectively. These levels of correlation between the two factors assist in evaluating potential distortions in the second factor due to the violation of linear independence among two items in the first factor. Since maximum likelihood estimation of model parameters was used in this study, all model estimates were calculated simultaneously. The prevailing opinion is that, with simultaneous estimation, misspecification in one part of the model may have an effect on estimates in other parts of the model (Kline, 2012). The different magnitudes of the

within-factor correlated error as well as different magnitude of between-factor correlations allows for the analysis of their potential for error propagation throughout the model.

Correlated error. Six levels of correlation between a single pair of within-factor error terms was evaluated (the error between the first and second indicators of the first factor). The levels of correlation represented a null value (no violation to independence assumption as a check on simulation performance) as well as a wide, plausible range of error correlation that has been observed in applied research as well as prior MC simulation research, including: 0.00, 0.10, 0.20, 0.30, 0.40, and 0.50. In the present study's model and generation of the data, the loadings were assigned unit variance along with the factors, by setting the error variances to vary around a value equal to one minus the square of the factor loadings. This allowed the loadings to be interpreted as standardized loadings. However, the error terms themselves are not standardized. Mplus allows the user to set a covariance. The error covariance would be equal to an error correlation if the errors had unit variance, which they do not. In order to set the error correlations to the desired values, the error covariance was set such that:

$$CORR(e_1, e_2) = \frac{COV(e_1,e_2)}{\sqrt{VAR(e_1)*VAR(e_2)}} \quad (2)$$

The error variances are determined by the factor loadings and, as previously stated, are equal to one minus the square of the loading. Thus for loadings equal to 0.70, the error variances equal 0.51, so a correlation of 0.50 would equal an error covariance of 0.255 (G. R. Hancock, personal communication, November 18, 2014).

In summary, in the population model, the sample sizes, loading magnitudes, between-factor correlations, and error correlation magnitudes were varied across conditions, but constrained to vary around a specific value within each condition. Additionally, the variances of

the latent factors were fixed equal to 1, and the residual variance, or error term, for each observed variable was specified to vary around a specific value, set according to the magnitude of its factor loading. Thus, the data are generated varying around the specified values according to model conditions.

The generated data were then analyzed using the same model without the correlated error specification (i.e., to determine the impact of ignoring the correlated error term). The factor variances in the analyzed model were set to 1 in order to scale the loading estimates, but the factor loadings and error variances were freely estimated using the same values used to generate the data as start values, instead of the default value of 1 used by Mplus.

Number of conditions and replications per condition. Each level of the four experimental variables described above was fully crossed, resulting in a total of 360 experimental conditions. Crossing all conditions allows for analysis of both systematic main effects as well as interactions among variables. For each of the model conditions, data was generated and analyzed 100,000 times. The resulting estimates for each parameter were then averaged across 100,000 replications per experimental condition, providing an average estimate for each of the 360 conditions to be analyzed.

RNG seed. The seed for the random number generator (RNG) used to generate the data from the model for each of the experimental conditions was set by the researcher as a combination of the sample size, between-factor correlation magnitude, loading magnitude, and correlated error magnitude for each set of replicates. For example, the condition in which the sample size was $N = 160$, the factor correlation = 0, the loading magnitude = 0.40, and the correlated error magnitude = 0, had a random number seed of 160040. These seeds were

generated in this manner instead of being freely varied in order to ensure that data could be regenerated later, if necessary (Bandalos & Leite, 2013).

Results Saving

Multiple pieces of data were collected for each of the simulations, including parameter point and interval estimates, *p*-values, and model fit statistics. Model fit statistics collected included χ^2, RMSEA, and SRMR. Summary statistics, provided by Mplus, for each combination of conditions, across replicates that converged, were pulled from the output and aggregated for analysis. The main variables of interest captured include the population values for each of the manipulated conditions, as well as the average estimate across replicates of each factor loading. Raw bias, the difference between the population value and the model estimated value, was calculated. Relative bias, which is the difference between the estimated and population value divided by the population value, giving a percent difference between the estimated model and population values, was also calculated. Coverage was measured by the percentage of 95% confidence intervals containing the parameter for each of the loadings and correlations. Finally, the percentage of replications converging was recorded as well.

Analytic Plan

Initially, the means for each of the 360 experimental conditions were computed and saved. Tables of means by experimental condition were created, as were figures of trends across levels of the factors. For the null and non-null conditions, separate analyses of variance (ANOVAs) were conducted to test main effects and two-way interactions among the experimental conditions on each of the key outcomes. Any main effects were then followed up with post-hoc contrasts, with alpha adjusted for multiple comparisons.

Null conditions. First, in order to test that the MC CFA simulations were functioning properly, a 3-factor, main-effects only analysis of variance (ANOVA) was conducted for all simulations in which the error correlation was set to 0 no bias due to a violation of linear independence. The three main effects tested included sample size, loading magnitude, and between-factor correlation magnitude. Separate ANOVAs were conducted on relative bias and coverage for the first variable's (Indicator X1) loading onto the first factor (Factor 1) (see again Figure 3 for illustration of X1's location in the model). In total, there were sixty null conditions with the correlated error magnitude set to 0. Any significant main effects for these models were followed up with Tukey's HSD multiple comparison procedure (MCP). Tukey's test compares all possible pairs of means while both adjusting for Type I error inflation as well as preserving maximum power.

All conditions. After establishing null results, 4-factor ANOVAs with all 2-way interactions were conducted on means from all of the fully crossed conditions (using all possible higher order interactions is not feasible due to a lack of degrees of freedom when each cell is its own condition, as it is in this case). The main effects included sample size, loading magnitude, between-factor correlation magnitude, and error correlation magnitude. Estimates for multiple variables (not just the first indicator of the first factor) were used as outcomes, and are further enumerated later. The dependent variables for each of the ANOVAs were selected based on prior research and potential results. Each variable selected represented a unique relationship with the correlated error term, and thus a potential for a different effect. Three separate relationships are found within the model. In relation to the correlated error term, the

indicator can be: within the same factor and directly affected, within the same factor and not directly affected, or in a different factor altogether.

Since the latent factors were scaled by setting the variance of each factor equal to 1, the loadings for each indicator was freely estimated, leaving the estimated loadings for each indicator a viable potential dependent variable. Indicator X1 (the first indicator loading onto the first factor; see Figure 3) was chosen to represent an indicator that is directly affected by the correlated error term. Since indicators X1 and X2 (the second indicator loading onto the first factor) were set to have shared error variance, the effect is likely to be similar, so only one is needed for evaluation. To evaluate the effect of the correlated error term on indicators without error covariance, but within the same factor as the items sharing error covariance, indicator X3 was chosen. Again, a similar effect is expected for both X3 and X4 (third and fourth indicators loading onto the first factor, with independent errors), so only one was chosen for analysis. Finally, in order to test the potential of error propagation to the loadings of indicators on a different, but correlated factor, indicator X5 (first factor loading onto the second factor) was used in analysis.

For each of the three indicators chosen for analysis, two different dependent variables were used. Relative bias for each of the indicators was as the primary outcome of interest. Relative bias takes the magnitude of the population estimate into account, allowing for better comparisons across indicator loading magnitude levels. Additionally, 95% confidence interval coverage was used as the dependent variable. Evaluating the effect on of the misspecification on confidence interval coverage provides information on several aspects of the model. Confidence interval coverage can be affected by both bias in the estimate and bias in the

standard errors. The evaluation of the change in coverage across different conditions allows the examination of not only how bias may be effecting parameter estimates, but also gives an idea of how standard errors are being effect by any bias. At the same time, 95% confidence interval coverage also provides insight into the effects on Type I error rates. Consequently, using 95% confidence interval coverage provides information on multiple aspects of the model, while limiting the number of tests needed, thereby limiting the necessity of Type I error adjustments.

As a precaution against Type I error inflation due to multiple tests on the same sample of data, the significance level for each ANOVA on the non-null conditions was adjusted using the Bonferroni (Dunn) procedure which divides the alpha level by the number of tests; consequently, based on six ANOVAs, the alpha level was adjusted to 0.0083.

MCPs. Trend analysis was identified as a useful orthogonal multiple comparison procedure (MCP) for evaluating the functional form of the relationship between levels of each main effect and each of the outcomes. Trend analysis can be thought of as a combination of orthogonal contrasts and polynomial regression (Lomax, 2007). Essentially, trend analysis is a set of orthogonal contrasts formed between sets of trends, where each subsequent equation is raised by a higher power. The initial trend is linear, followed by quadratic, cubic, and higher if necessary. The significance for each contrast is tested with a traditional *t*-statistic, with the contrast estimate denoted with ψ.

Chapter IV: Monte Carlo Simulation Study Results

The first step in evaluating the Monte Carlo simulation results was the inspection of model convergence statistics across conditions. Overall, 98% of all models converged successfully. Conditions involving the smallest sample size, $N = 80$, showed a slightly lower

convergence rate, 96%, than did the larger sample size conditions, *N* = 160 and 400, which had average convergences of 99% and 100%, respectively. Conditions involving the smaller loading magnitude of 0.40 also showed lower convergence rates, averaging 93%, whereas the other loading magnitudes converged at 99% of the time or better. Indeed, the conditions with the worst convergence rates involved models with small sample sizes, *N* = 80, and small loading magnitudes, 0.40, which resulted in an average convergence rate of 82%.

Having evaluated model convergence, multiple indicators of the importance of each result were necessary to evaluate the results of each test of significance. While the significance of each result, judging its likelihood of stemming from chance, was of primary importance, the nature of the large number of replications used in this study called for expanding the criteria used to determine the practical importance of each result. For all tests of significance in this study, the error terms were exceedingly small. The amount of variance accounted for by the predictors in each of the ANOVAs, R^2 = 1.00 to 0.976 and Adj. R^2 = 0.999 to 0.969, left little variability for the error terms. This, in turn, led to the significance of many tests for which practical importance to applied research could be called into question. In order to properly address and interpret the relative importance of the significant effects within the ANOVAs, the effect size, ω^2, was used. ω^2 is determined as:

$$\omega^2 = \frac{SS_{betw} - (J-1)MS_{with}}{SS_{total} + MS_{with}}. \quad (3)$$

The ω^2 value was chosen in this instance as it is known to be less biased than the commonly used η^2 (Lomax, 2007). Eta-squared is a simple proportion of the amount of variation in the outcome that is explained by the predictors. According to Cohen (1988), ω^2 = 0.01, 0.06, and 0.14 are interpreted as small, medium, and large effect sizes, respectively. The mean difference

in bias was also used as an indicator of effect size. Muthén, Kaplan, and Hollis (1987) defined a relative bias of 0.10 as a good rule of thumb for bias that might affect model interpretation. Other authors have recommended less stringent standards for considering an estimate biased (Hoogland & Boomsma, 1998), however 0.10 appears to be a good basis for comparison in the current study. The trend analyses, however, do not lend themselves to simple measures of effect size. To help evaluate the relative importance of significant results in the trend analyses, plots of means were generated and evaluated (Lomax, 2007). Visual evidence of the polynomial relationship of the variables should be present in the mean plots if the significant results of the trend analysis are indeed meaningful in an applied setting.

Descriptive Statistics

The initial inspection of the means provided some insight into not only potential bias in the estimated loadings, but also the patterns of bias. As mentioned in the Data Analysis Plan (Chapter III), for each loading, means were computed for the loading point estimate, 95% CI coverage, Raw Bias, and Relative Bias for each of the 360 experimental conditions. Recall that 95% CI coverage indicates the number of 95% CI calculated that contain the true parameter location, while raw bias is the difference between the true and estimated values and relative bias is the difference between the true and estimated values divided by the true value, yielding a percentage of bias based on the magnitude of the parameter. Though the loading point estimates and raw bias were not included in the further analyses of the data, their descriptive information is provided for reader interest. The loading estimates give an idea of how the average loading compares to the known value. For instance, across all conditions, the average

loading should equal 0.60; deviation from this indicates bias, which is especially easily seen in the raw bias estimate.

As shown in Table 1, both the loading estimates and raw bias indicate some deviance from the true values. The pattern of the bias also appears to follow that found by Cole et al. (2007), with indicators affected by the error covariance appearing to have positively biased loading estimates, while those within the same factor, but unaffected by error covariance, having negatively biased loadings. This pattern appears to continue even when the means are disaggregated by sample size, loading magnitude, between-factor correlation magnitude, and error correlation magnitude, as seen in Table 2, 3, 4, and 5, respectively.

The 95% Confidence Interval (CI) coverage for Indicators X1 and X3 also appears to have a discernable pattern, worthy of further analysis. The coverage for the 95% CI surrounding the estimated loadings for Indicators X1 and X2 are well below their true value across all main effects, excluding null conditions, as seen in Table 1 through 5. The results shown in Table 2 indicate that as the sample size increases, the coverage of the 95% CI for Indicators X1 and X3 decrease. This contrasts with Indicator X5, where the coverage appears to improve slightly as the sample size increases.

Overall, the means, broken down by the main effects, appear to show some effect of the under-specification of correlated error terms. There are observable differences in loading bias and CI coverage among levels within each of the factors, excluding null conditions. The size of these differences vary throughout, but are certainly worth further analysis.

Null Condition Results

For null condition models with no correlated error term, the average model convergence rate was 98%. The results of the 3-factor main-effects only ANOVA for the null condition on relative bias showed a significant main effect for loading magnitude for X1 on Factor 1, $p < 0.001$, $\omega^2 = 0.460$ (none of the other main effects were significant (ps > 0.05)). The significant loading magnitude effect, which indicates that the amount of relative bias in Indicator X1 differs over different levels of loading magnitude, was further evaluated using Tukey's HSD. Tukey's HSD showed significant differences between the loading magnitude equal to 0.40 and each of other loadings, 0.50, 0.60, 0.70, and 0.80, ps < 0.001. The mean differences between 0.40 and the other levels ranged from 0.017 to 0.020, indicating a small difference in the levels of relative bias in the null condition for loading magnitude = 0.40. While the level of relative bias in the 0.40 loadings condition is quite small, it is worrisome that small levels of bias are induced for this condition when zero error correlation is present. Recalling that loading magnitudes were held constant across all indicators for a given model for this study, the bias in the lowest level of loading magnitude is likely due to instability in the estimates when all of the loadings are small (Bentler & Yuan, 1999). Still, the results warrant further evaluation in future research. Despite this, the differences between all other loading magnitude levels were found to be non-significant, indicating no differences in relative bias based on the loading size for larger loadings. Overall, the results for the evaluation of conditions with no error correlation indicate the model is functioning properly, with small deviations for the loading magnitude of 0.40.

In terms of overall coverage for null conditions, the results showed that the mean 95% CI coverage for X1's loading was 94.35% (*SD* = 0.50%), which indicates a very slight underestimate of coverage across all models.

The 3-factor ANOVA on coverage again showed a significant main effect for loading magnitude, $p < 0.001$, $\omega^2 = 0.267$, but in addition, also showed an effect due to sample size, $p < 0.001$, $\omega^2 = 0.577$. No significant differences were found over levels of between-factor correlation magnitude, $p > 0.05$. Tukey HSD tests indicated significant differences in coverage between loading magnitude 0.40 and all higher level magnitudes were significant (as found before), $p < 0.001$, as well as between 0.50 and higher loading magnitudes, ps < 0.05. The mean differences in CI coverage between levels of loading magnitude for significant results were quite slight, ranging from 0.24% coverage difference between 0.50 and 0.60 conditions, to 0.68% coverage difference between 0.40 and 0.80 loading magnitude conditions (*SE* = 0.10%). Importantly, these differences in coverage due to loading magnitude are likely the result of (a) a small bias in loading estimates observed for the .40 loading condition (already described above), and (b) a small bias in the estimate of the standard error, particularly for lower indicators with lower factor correlations.

All levels of sample size were significantly different from each other on coverage, ps < 0.001; in general, as sample size was increased, the coverage improved. This said, these differences were substantively quite small: coverage differences ranged from 0.32% between $N = 160$ and $N = 400$ conditions, to 0.93% between $N = 80$ and $N = 400$ conditions (*SE* = 0.07%).

Given the slight underestimation of coverage across all models and particularly those conditions with both lower loading magnitude levels and smaller sample sizes, confidence interval results for the non-null conditions (forthcoming) should be interpreted with these null results in mind (of approximately 0.5% to 1% lower coverage than expected).

All Condition Results

Bias. As shown in Table 6, the results of the ANOVAs on relative bias across all study conditions showed significant main effects for sample size, loading magnitude, between-factor correlation magnitude, and error correlation magnitude in Indicators X1 and X3. However, because the estimates are so precise (100,000 replications per condition were used), not every statistically significant effect may be "practically significant." Hence, examination of effect size provides deeper insight into the relative importance of the significant findings. Both sample size and between-factor correlation main effects show negligible effect sizes on bias for X1 and X3 (ω^2s < 0.01), while loading and error correlation had large effect sizes (ω^2s ranging from 0.324 to 0.426 for loading effects, and 0.398 to 0.509 for error correlation effects). This indicates that not only does the failure to specify correlated error terms cause bias in loading estimates within a confirmatory factor model, but the main drivers of the relative magnitude of this bias are the magnitude of the factor loading along with the magnitude of the correlated error term.

Plots of mean bias by condition given in Figure 4 and 5 illustrate the effects of the misspecification on Indicators X1 and X3, respectively. When the indicator shares error covariance with another indicator (X1), positive bias increases as the magnitude of the correlated error increases. For indicators within the same factor, but unaffected by error covariance (X3), the bias is instead negatively increasing. The results of a trend analysis on the

relationship between loading magnitude and relative bias indeed showed significant positive linear effects on bias for X1, with $\psi_{linear} = 0.375$, $p < 0.001$, and significant negative linear effect for X3, $\psi_{linear} = -0.213$, $p < 0.001$. Although the cubic trend was also found to be significant for each, which would indicate two bends in the functional form of the relationship between error covariance and bias, this is likely due to the ceiling effect on relative bias.

Similarly, the results of a trend analysis on relative bias by levels of loading magnitude indicated that the linear trend was significant for X1, $\psi_{linear} = -0.340$, $p < 0.001$, as well as X3, $\psi_{linear} = 0.153$, $p < 0.001$. The quadratic trend was also significant for both indicators (ps < .001), with X1 exhibiting a positive quadratic and X3 exhibiting a negative quadratic. Additionally, for X1, loading magnitude had additional higher order cubic and quartic trends (likely due to precision of estimates and ceiling effect). In sum, for indicators affected by error covariance (X1), bias decreases as magnitude increases, but this decrease slows down as it reaches the maximum magnitude threshold. For indicators unaffected by shared error but which load on the same factor as indicators with the problem (X3), the bias increases (albeit at a lower rate) as magnitude of loading increases, but then decelerates as the loading magnitude reaches its maximum value.

Finally, as can also be seen in Figure 4 and 5, the interaction effect between loading magnitude and error correlation magnitude on relative bias is ordinal, and shows that effect of the error correlation on bias in both X1 (which is positively biased) and X3 (negatively biased) is mitigated by increased loading magnitudes. In other words, the weaker the connection between the indicator and the latent factor, combined with a greater error covariance between indicators, can lead to larger amounts of relative bias.

All other interaction effects observed in Table 6, although statistically significant, had effect sizes below the threshold to even be considered “small”. Accordingly, they are not interpreted or analyzed further.

Coverage. Surprisingly, the effect of sample size on relative bias, though statistically significant, had a negligible effect size on relative bias. This said, the effect sizes for 95% CI Coverage were substantial (see Table 7 for ANOVA results).

Recall that 95% Confidence Interval Coverage is computed using both the estimate of the loading as well as the standard error of that estimate. Analysis of the relative bias in the estimates showed that both loading magnitude and error correlation had deleterious effects but that sample size did not meaningfully affect the amount of bias in the loading estimate. However, the standard error of the estimates do appear to have been affected by sample size. As shown in Table 7, the 95% CI Coverage in X1 and X3 were affected by sample size, loading magnitude, and error correlation (*p*s < 0.001), each with medium to large effect sizes ranging from $\omega^2 = 0.095$ to 0.974. The effect sizes on 95% CI coverage for X1 were all greater than for X3. Consistent with the results for relative bias, the main effect of between-factor correlation magnitude was insubstantial in size for both indicators.

Visual inspection of the means plotted in Figure 6 and 7 for X1 and X3, respectively, shows that the larger the error correlation, the worse the CI coverage becomes. High levels of error covariance can result in 95% Confidence Intervals capturing the population value of X1’s loading in less than 60% of the intervals for very small sample sizes, and less than 10% of the intervals for relatively large samples sizes. In other words, the contribution of the previous bias in the point estimate of the loading due to correlated error (but not sample size) to 95% CI

coverage is further exacerbated by bias due to increasing sample size. This latter finding indicates that the standard error for the X1 loading will become increasingly downwardly biased as both correlated error and sample size increase. This finding is also true for X3's coverage, albeit to a lesser extent.

The follow-up trend analyses showed that linear effects of sample size on coverage for both X1 and X3 were significant and negative, with $\psi_{linear} = -0.207$, $p < 0.001$, and $\psi_{linear} = -0.118$, $p < 0.001$, respectively. As seen in Figure 6 and 7, the decrease in coverage for X1, the variable affected by error covariance, is nearly twice as much as that for X3, the variable unaffected by error covariance but which loads on the same factor as X1. The quadratic term (deceleration in negative bias) was also significant and similar in value for both, $\psi_{quadratic} = -0.028$, $p < 0.001$, and $\psi_{quadratic} = -0.030$, $p < 0.001$. In other words, a slight deceleration in the problematic coverage occurs as coverage moves towards a floor effect of 0.

Trend analyses of the other main effect conditions on coverage showed expected patterns. Specifically, there were negative linear and quadratic effects of error correlation levels on coverage in both X1 and X3, but the linear effects were again much greater for X1. For X1, $\psi_{linear} = -0.545$, $p < 0.001$, and $\psi_{quadratic} = -0.098$, $p < 0.001$; for X3, $\psi_{linear} = -0.312$, $p < 0.001$, and $\psi_{quadratic} = -0.086$, $p < 0.001$. Although small cubic trend effects were found for both indicators, this was likely due to precision in coverage and floor effects.

The CI coverage also showed a distinct trend across the levels of loading magnitude, though in a positive direction. For X1, $\psi_{linear} = 0.185$, $p < 0.001$, and $\psi_{quadratic} = 0.072$, $p < 0.001$; for X3, $\psi_{linear} = 0.128$, $p < 0.001$, and $\psi_{quadratic} = 0.010$, $p < 0.01$. For X1, the bias in coverage in largely unchanged between the lowest loading levels, seen in Figure 10, then climbs at an

increasingly faster rate as the loading value increases. A similar pattern is found for Indicator X3, although the coverage does increase between the lower levels. Though larger sample sizes appear to have a detrimental effect on CI coverage with the failure to specify a correlated error term, larger factor loadings, representing stronger connections between the observed and latent variables, improve the CI coverage. The mean CI coverage for both indicators starts well below 95% at the lowest loading levels and never reaches its proper levels even on loadings of 0.8.

In summary, for the relatively simple CFA model chosen for this study, the impact of the main effects on the bias created in both loading estimates and 95% CI coverage follow an identifiable and potentially calculable trend. In general, the bias tends to be greater and increase or decrease more sharply for Indicator X1 than for Indicator X3. This indicates that the failure to specify the correlated error term has a greater effect on the indicators sharing error covariance than those simply within the same factor. When compared to the patterns of relative bias among the main effects, it appears likely that the CI coverage is affected by not only the bias in the loading estimates, but also by the attenuation of the standard errors, exacerbating the bias in the CIs. While the patters of bias for different levels of each main effect are rather simple when averaged across the model conditions, the overall complexity of the model in general, with multiple predictors of the bias, makes calculating the exact effect of the under-specification given specific model conditions a difficult undertaking.

Indicator X5. Recall that the indicator X5 was the variable loading onto Factor 2 that had no error correlation with any other variable in the model (see again Figure 3 for illustration of location in the model). The patterns of significance for this indicator's relative bias (Table 7) and

95% CI coverage (Table 8) were similar. The main effects for sample size, loading magnitude, and between-factor correlation magnitude were significant (*p*s < 0.001), as were the 2-way interactions among these variables. As seen in Table 6, relative bias in X5 shows large effect size only for loading magnitude ($\omega^2 = 0.477$) as well as its interaction with sample size ($\omega^2 = 0.422$); the other two main effects are negligible in size. For coverage, seen in Table 7, the results are nearly identical, with large effect sizes for sample size ($\omega^2 = 0.594$), loading magnitude ($\omega^2 = 0.270$) and their interaction ($\omega^2 = 0.082$).

Plots of the bias and coverage by loading magnitude and sample size for X5 are given in Figure 8 and 9; it is clear from these illustrations that the main effects and interaction are driven by the instability in the estimates for X5 when the sample size is small, $N = 80$, and the factor loadings are small, 0.40, as was already observed in the analysis results for the null condition with no correlated error between observed variables.

The results of the trend analyses on relative bias and coverage for X5 showed that sample size and loading magnitude had mostly off-setting linear and quadratic effects on bias, with sample size's trends of $\psi_{linear} = -0.002$, $p < 0.001$ and $\psi_{quadratic} = 0.002$, $p < 0.001$, and loading magnitude's trends of $\psi_{linear} = -0.015$, $p < 0.001$ and $\psi_{quadratic} = 0.011$, $p < 0.001$. The between-factor correlation, however, showed only negative effects on bias, with $\psi_{linear} = -0.003$, $p < 0.001$ and $\psi_{quadratic} = -0.001$, $p < 0.001$. Not surprisingly, trend analyses on coverage showed the same pattern in results.

Most importantly, the main effect for error correlation and its associated 2-way interactions on bias and coverage for X5 were not significant. Overall, the results for X5 indicate

that bias due to error correlation does not propagate across the model from one factor with affected items to the next factor with no affected items.

Summary. The magnitudes of error correlation and loadings had an important effect on the amount of bias created by the under-specification of the correlated error term. For items within the factor effected by the under-specification, the larger the error correlation, the larger the relative bias, while larger loadings induced smaller relative bias. Also clear was the differential effect of error correlation magnitude by loading magnitude. These two model conditions had competing effects on the amount of relative bias created. Though statistically significant, sample size and between-factor correlation did not affect the relative bias in a meaningful way.

Coverage of the 95% CI for each of the model estimates under investigation suggested that both the bias in the estimates, as well as attenuation of the standard errors, led to lower levels of coverage for certain conditions. As expected, lower loading magnitudes, coupled with increased error correlation, significantly lowered CI coverage due to the already present bias in the point estimates. However, these effects were amplified by increasing sample size, which in turn must have had a shrinking effect on the loading variance estimates since sample size did not play any role in biased point estimates.

Finally, despite the potential for error to propagate throughout the model (Kline, 2012) there was no indication of this effect in the model used in the present study. The loading estimates on the factor with unaffected items showed no bias due to the under-specification of correlated error terms. Within this relatively standard model, error covariance among indicators on one factor does not affect loading estimates on another factor.

Chapter V: Applied Analysis Demonstration

Following the Monte Carlo (MC) simulation study, the same principles of the analysis were applied to an extant data source based on recent educational research survey data to demonstrate the difference between specifying a model that ignores linear dependence and one that directly models the dependence as a correlated error term.

Sample

The data for this analysis comes from a multi-institutional study examining the links between multiple forms of belonging and engagement in undergraduate students (Wilson, Jones, Bocell, Kim, Veilleux, Floyd-Smith, Bates, and Plett, In Press). While the original study evaluated students at five different higher education institutions, only one institution was used for analysis in the current study. The sample used in the current analysis consisted of sophomores, juniors and seniors enrolled in science, technology and math courses at a major research university in the northwestern United States. Student participation was voluntary, and completed on both paper-and-pencil and electronic versions of the survey. The total sample size was *N* = 886 participants, with 599 males and 278 females. The sample included 305 seniors, 391 juniors and 190 sophomores. 37% self-identified as Asian/Asian American, 1% as African American/Black, 47% as White/Caucasian, 2% as Hispanic and 11% as other. Although the sample size is more than double the largest sample in the present MC simulation study, the sample from other colleges originally included in the data set ranged from *N* = 63 to 274 per college, which is more in line with the samples evaluated in the simulation. The institution for the current study was chosen by necessity, due to accessibility to the data. Despite the large

sample size, the MC study results are applicable to this larger data set, as the effect of sample size on relative bias was found to be non-significant.

Measures

Two scales from the larger survey were used for the present analysis. The latent variables measured were grounded in psychological theory, and termed “Belonging to Major” and “Belonging to Class”. The specific items indicating each latent variable were adapted from the Belonging Scale by Anderson-Butcher and Conroy (2002). They were meant to assess a student’s feelings of support and acceptance in the STEM class the survey was conducted in and in their major. Each of the scales consisted of four items using a 5-point Likert Scale, ranging from strongly disagree to strongly agree. The original research on the scale found strong internal consistency ($\alpha = 0.93$) and used CFA to establish the factorial validity of the scales. In this study, acceptable levels of internal consistency in the sample were found for both Belonging to Major ($\alpha = 0.834$) and Belonging to Class ($\alpha = 0.878$).

Belonging to Major consisted of four items, labeled A01, A16, A20, and B09. Belonging to Class comprised four items as well, labeled I09, I15, I18, and I20. The letter for each items indicates the section of the survey the item was in, while the number indicates the item’s location within the section. The items for Belonging to Major included language such as “I feel that I am a part of this major” (A01), and “I feel that I am accepted in this major” (B09). The language for items in Belonging to Class included, “I feel that I am supported in this class” (I09) as well as “I feel that I am a part of this class” (I15).

These two latent variables were chosen for the present analysis for several reasons. First, having four indicators each, they closely matched the model under investigation in the MC

simulation study. Second, previous analysis of the scales for a separate study indicated the potential for error covariance among items within the Belonging to Major factor. In the prior study, one item was removed from the scale because of the potential violation of the assumption of linear independence. However, the item was included in the present analysis to help determine the impact of failing to account for such violations through the under-specification of correlated error terms. Modification indices were computed and the expected improvement in model fit for the addition of correlated error terms between observed variables was evaluated to aid in the decision of where to specify the correlated error term. As noted earlier, modification indices (MIs) give the expected χ^2 change if a path currently confined to zero is freed and allowed to very. A large MI indicates freeing the estimate could have a relatively large impact on model fit.

Analytic Plan

The analysis took place in two stages. The first stage involved the specification and estimation of two separate, though nested, CFA models. As with the simulation study, the CFA models were analyzed in Mplus 7.2 (Muthén & Muthén, 1998-2012). Two correlated factors, each with four indicators, were specified. By default, Mplus sets the loading of the first indicator for each latent factor equal to 1 as a metric to scale the other estimates within the model, this is a common method for scaling the latent variable (Bollen, 2002). Standardized model estimates, based on specifying factor variances equal to 1, were also calculated, for comparison to the standardized values used in the MC simulation study.

To estimate the effect of failing to account for error covariance between items first the model was estimated failing to specify a correlated error term between any items. The

assumption of linear independence was presumed tenable. Both standardized and unstandardized model estimates were recorded, along with model fit statistics. The χ^2, RMSEA, SRMR, CFI, and TLI were all used to evaluated model fit. A discussion of each is given below.

- The χ^2 statistic indicates the deviation of the fitted model from the sample covariance matrix. Research has shown that χ^2 is sensitive to sample size and even small deviation in fit would show a significant difference in large sample sizes (Hu & Bentler, 1995).
- The standard approach for evaluating the root mean square error of approximation (RMSEA) has been to consider values less than or equal to 0.05 as a close approximate fit. Values between 0.05 and 0.08 are often considered a reasonable fit, with larger estimates indicating poor fit (Browne & Cudeck, 1993).
- The square root of the mean squared residual (SRMR) is a measure of the ability of the model implied correlation matrix to reproduce the observed correlation matrix. A general cutoff value of 0.09 for SRMR was recommended by Hu and Bentler (1999).
- The values indicating good fit using the comparative fit index (CFI) and other incremental fit indices, such as the Tucker-Lewis index (TLI), are not widely agreed upon. Kline (2005) cited 0.90 or higher as indicative of good fit, while Hu and Bentler (1999) recommend values of 0.95 or higher should be used as the cutoff in some circumstances. Though the fit of the data to the model is not of primary interest, model fit should be inspected and found to be reasonable before model estimates are inspected and interpreted (Hancock & French, 2006).

All of these modification indices were evaluated to help inform the specification of the correlated error terms between items. The correlated error term was limited to a single within-

factor correlation between a single pair of items in order to best demonstrate the specification of the general 2-factor model used for the MC simulation study.

Once the items were selected for the correlated error term, the previous model was analyzed with a correlated error term between two items freely estimated. Both standardized and unstandardized loadings were recorded. Raw and Relative bias was calculated for all factor loadings, using the model with the correlated error term as the reference. In this circumstance, the model with the correlated error term was assumed to be the correctly specified model, with the other model representing an under-specification. This approach again mirrors the MC simulation study, providing a direct comparison. Comparisons were then made to the results of the MC study to compare the effects of potentially under-specified error correlations. Appendix IV provides general Mplus code for the applied analysis.

Results

Due to the use of modification indices, the inclusion of the correlated error term provided a small but noticeable improvement in model fit for some of the fit indices (see Table 8). Both models indicated reasonable to good fit for all of these indices, excluding the χ^2 test of model fit. Although both χ^2 values were significant (indicating a poor fit of the model to the data), the correctly specified model that included the error correlation term did show a significantly better fit to the data, $\chi^2_{change(1)} = 29.541$, $p < .001$. In addition, the RMSEA estimates were slightly higher than what would be desired for the incorrect, under-specified model, with the correctly specified model showing a markedly smaller estimate. However, neither would be interpreted as displaying a “close approximate fit” fit (recall that a “close approximate fit” is an RMSEA ≤ 0.05), instead they show “reasonable” fit (RMSEA = 0.50 to 0.08). All of the other fit

indices, including SRMR, CFI, and TLI, indicated good overall model fit for both models, but with the correctly specified model (the one with the error term specified) showing a slightly improved fit. These results appear to be in keeping with prior research showing that correlated error misspecifications did not generally affect model fit, so this lack of difference between models is not surprising (Heene et al., 2012).

Having determined the general acceptability of the fit of the model to the data, further interpretation of the model estimates is appropriate. As with the MC simulation study, the focus here lies in the comparison of the estimated loading value between the correctly specified and under-specified models. Selected results for the analysis can be found in Table 9.

First and most noteworthy, the test of the standardized estimate for the error correlation between Items A01 and B09 (the two correlated items in the first factor) revealed a significant amount of covariance between the error terms. The estimate corresponds to the middle range of the error correlations evaluated in the MC simulation study, at 0.279. Under the assumption that this represents a true error covariance between the construct irrelevant variance in the observed variables, the previous results would indicate the potential for bias in the loadings given an under-specified model failing to account for the error covariance.

The MC simulation study results showed a pattern of bias where items affected by the error covariance displayed positive bias, while items within the same factor, but unaffected by the error covariance, displayed negative bias. This pattern directly matches the results of the model estimates in the applied analysis, seen in Table 9, in which the correlated error term was not specified. For this model, the loadings for the two items specified with a correlated error term, B09 and A01, were positively biased by an average of 8% bias in the under-specified

model, and the items within the same factor, but unaffected by the error covariance, exhibited negative bias of approximately 4%. Although the bias in B09 and A01 was below the 10% threshold set by Muthén et. al. (1987), the bias was above the threshold of 5% recommended by Hoogland and Boomsma (1998), in some instances.

Recall that presence of attenuated 95% confidence intervals (CI) can be derived from both the point estimates as well as their standard errors. If the CI coverage is attenuated in the model without the correlated error terms specified, then either the loading estimate and/or its standard error would be smaller than in the model with the correlated error term specified. The results of the applied analysis follow this prediction. As seen in Table 9, the estimates of the standard errors for the first of the two of the observed variables, B09 and A01, with the correlated error term in the first factor was smaller in the model without the correlated error term specified (a decrease from .019 to .015 in the standard error estimate). Meanwhile, the same pattern was found for the first of the two unaffected variables in that same factor, A20: its standard error decreased from .025 to .019 in the model without the proper correlated error specification. Hence, both relative bias and coverage may be negatively affected in applied research that does not account for correlated error terms.

Finally, as was observed in the MC simulation results, the bias from the omitted correlated error term in one factor did not propagate to the other factor's item loading estimates or their standard errors. The estimated standardized covariance of 0.657 between the two factors themselves (Belonging to Class and Belonging to Major) represents a large, significant portion of shared variance. Importantly, despite the high covariation between the

factors, any bias created by failing to account for error covariation among items in one factor did not spread to the other factor.

Chapter VII: Discussion

Monte Carlo Simulation Study

There is little doubt, given the standard CFA model, that violations of linear independence among indicator items results in bias among the loading estimates and their corresponding 95% confidence intervals. Based on the large R^2 values for each of the ANOVAs conducted on the mean estimates, the major components determining the level of bias have likely been identified by the MC simulation.

As one might expect, the magnitude of the error correlation and factor loadings both play a significant role in determining the level of bias in the loadings and in the CI coverage, but in opposite directions. The higher the error correlation is, the greater the bias in loading estimates of affected observed variables and the worse the coverage in models that ignore the correlation; conversely, the higher the loading magnitude between an affected item and its factor, the lower the bias and better the coverage in models that ignore the violation to independence. Further, the interaction between correlated error and loading magnitude meant that bias was amplified by the joint presence of both small loadings and high error correlations. Another way to view this is that stronger associations between the indicators and the latent factor help to ameliorate the effect of violations of linear dependence among the items in that factor.

Importantly, although sample size was not an important indicator of bias in the loadings, its effect on coverage pointed to attenuation of affected loadings' standard errors. The effect of error correlation and loading magnitude on point estimate bias in affected variables was compounded by increased sample sizes.

Other researchers have indicated that effects from misspecification errors within one part of a structural equation model may spread, causing problems with estimates in other parts of the model (Kline, 2012). Depending on the perspective taken, this is both confirmed and refuted by the research at hand. On the one hand, all indicators within the factor with the two affected items (items in Factor 1) were subject to varying levels of bias depending on model conditions. As such, unaffected items within the factor were negatively affected and hence could constitute as a "spread" of a problem in one part of the model to another part of the model. However, on the other hand, estimates of item loadings and their coverages in the second factor, for which there was no violation, were unaffected by the problems in Factor 1; this held true across a variety of between-factor correlations. In this sense, the violation of linear independence did not propagate across the model.

The bias in the loading estimates may affect the interpretation of the model results. Indicators in violation of the assumption of linear independence may seem as though they are better indicators of the latent construct, when in fact they are no better than items within the same factor with no violation of the assumption. In addition, items maintaining linear independence within the same factor will then appear to be poorer indicators of the latent variable. The combination of these bias, in different directions, could skew the researcher's perceptions of an item's relationship to the latent factor, or even change the evaluation of the scale as a whole.

In models where the indicators do not correspond well with the latent construct and construct-irrelevant variance is share among items, the effect of the bias could be profound.

The positive bias in loading for items affected by error covariance could give a false impression as to the quality of the indicators. Low factor loadings are made specious, in addition to their violation of linear independence, possibly causing assessment builders to erroneously include them in scales. Under field test conditions, assessment builders may have a pool of items under consideration for the test or survey. Though CFA may only be a portion of the decision making process, biased loading estimates may give false impressions of the psychometric properties of the items and scales. Combined with other psychometric information, such as IRT estimates or g-studies, the picture may become confused, increasing the potential for including items better left of the assessment.

The wild variations in the coverage of the 95% confidence intervals also indicate the presence of inflated Type I error rates. Under the worst conditions, large error covariance, large sample sizes, and low loadings, 95% CI coverage can quickly drop, reaching 10% coverage for indictors with error covariance, and 35% for indicators on the same factor unaffected by the error covariance. Model estimates, in this case, will be given a greater chance at rejection than is normally assumed. Small estimated loadings will not only increase in magnitude, but they may be mistakenly significant, further increasing the chance of incorrect conclusions being drawn.

Generally speaking, as one would assume, the better the construction of the measurement tool, the less likely violations will have a meaningful impact on the estimates and interpretations. Strong relationships between the latent variable and its indicator items and low levels of error covariance among items invariably lead to better estimates. The value of the

model estimates appear to be directly related to the design and construction of the assessment they are based on.

Applied Analysis Demonstration

Model Fit. The evaluation of the effect of correlated errors on model fit statistics was not a part of the MC simulation study. However, as the evaluation of model fit is an integral part of model assessment and theory building, the judging of model fit was necessary for the applied portion of this study. Model fit indicates how well the model fits the data. Different indicators use different methods to evaluate the deviation of the model form the data. However, the fit indices for both the correct and under-specified models exhibited acceptable to good model fit. Under normal circumstances, both models would might have been accepted by the researcher with little further evaluation. Consequently, the final model would depend heavily on the initial specification, requiring a familiarity with method and context effects, and other sources of error covariance in order to specify what is termed the correctly specified model in this research.

Loading Estimates. The conditions of the applied analysis closely matched those in the MC study, providing added support for the usefulness of the MC results to applied data sets in general. Assuming the correctly specified model contains the correlated error term, loading estimates across both latent factors ranged from 0.664 to 0.845, within factors the range of loadings was smaller, as seen in Table 9. The similarity of loading magnitudes within and across factors shows the simplifying assumption of equal loadings within conditions may be feasible under certain circumstances. Additionally, the estimated error correlation of 0.279 falls in the middle of the range of correlated error terms tested in the MC simulation. With loadings of

0.70 and a correlated error term of 0.20 or 0.30, the levels of relative bias in the effected loadings, X1 and X2, in the misspecified model were equal to 0.072 and 0.118, respectively. The bias resulting from the misspecification in the applied model closely approximated these levels of relative bias, with estimates of 0.067 and 0.101 in B09 and A01, respectively. The estimated between-factor correlation of 0.657 in the correctly specified model was higher than the between-factor correlations modeled in the MC study. Still, there was no evidence that the bias propagated to loadings on the other factor.

Though the items did not exhibit a significant amount of bias according to the criteria set forth by B. O. Muthén et al. (1987), the combination of the positive and negative bias could change how the loadings are interpreted relative to each other. Taken alone, the positive bias in B09 and A01 does not drastically change possible interpretations of the loadings; both variables simply showed improvement to already high loadings. However, the combination of positive and negative bias within the factor increases the range of loadings in this instance, as well as changing the relative strength of the loading of each item when compared to other observed variables within the factor. A01 no longer has the lowest loading. Additionally, the negative bias in A20 makes its new position as the lowest loading more noticeable. This change may influence how the researcher evaluates and interprets the loadings of the observed variables.

Though small in terms of a validation study, the results of the applied analysis provide good evidence of the potential impact of the failure to specify correlated error terms within CFA. Not only does the failure to specify correlated error terms cause bias in the model estimates, but the conditions necessary to create the bias can be easily found in applied

research. While there is always the possibility of a researcher specifying correlated error terms to improve model fit under the auspices of taking into account the potential for error covariance due to the like of method effects. Certainly researches will stand on the divide between potential violations of linear independence and replicability of specification searchers. Perhaps there is no easy answer to the question of when the inclusion of correlated error terms is justified given the construction of the assessment or various aspects affecting a person's response.

Implications

Structural equation modeling is a useful, practical tool for test and survey scale development (Raykov, 2012). Large-scale assessments used for research, policy, and other important decisions are most likely to go through a rigorous scale development program which would include the use of factor analysis (DeVellis, 2012).

Generally the first steps include defining your variable of interest, creating item writing rules and building a pool of items to potentially include in the assessment. After a pool of potential items for the assessment is created and field tested on a sample of respondents, the pool of items needs to be narrowed down for use on the final assessment. There are many methods for selecting items to include or exclude for the final assessment, CTT and IRT for example. In both CTT and IRT, items with low discrimination, the ability to differentiate between students of varying ability, may be removed. Poor fit to the chosen IRT model may also cause the removal of items. A confirmatory factor analysis is also a commonly used as one piece of evidence to decide on item inclusion or exclusion, as seen in Kenney, Lac, Hummer, and LaBrie (2014). Items showing evidence of loading on more than one factor or sharing error

variance with another item are sometimes removed from the model. A low factor loading, signifying a week connection between the latent and observed variables is also used as cause to remove items from an assessment or survey.

The decision to include or exclude an item generally lies in the contribution the item makes to the validity and reliability of the assessment as a whole. Validity is a widely debated concept throughout the measurement community. On the whole, validity studies try to determine how well the latent variable measured by the tool corresponds to the latent variable of interest. As stated earlier, confirmatory factor analysis is one tool used to assess and provide evidence for validity, especially construct validity (Allen & Yen, 1979). There are many examples throughout the literature, where a CFA is used as one of the primary indicators of the validity of a scale (See Bostic, McGartland Rubio, & Hood, 2000; Mack et al., 2015). The strength of the association between the indicators and the latent factor (the loadings) represents how well the indicators represent the latent factor of interest. As seen in the current paper's Monte Carlo study and applied analysis, violations of linear independence can increase the loadings of some indicators, while decreasing others. This may distort the validity of an assessment, especially when the analysis is used as primary evidence towards the construct validity of the measure.

In order to avoid situations where model estimates are biased and the resulting validity evaluation is questionable, care must be taken in two specific phases of assessment and survey construction and evaluation. The first phase is item writing. Since the bias is most pronounced at lower loading magnitudes, sound item writing practices would help to strengthen the likelihood of high loadings, and thus low bias, when linear independence is violated. Well researched recommendations for item writing are available for surveys (Dillman et al., 2009)

and educational assessments (Taylor & Nolen, 2005). In addition to helping ensure strong loadings, following item writing recommendations would also reduce the possibility of error covariance among items assumed to hold linear independence. A well written assessment or survey will go a long way towards avoiding bias caused by a violation of linear independence. The second phase is during model specification. Here, the research cited earlier in Chapter I regarding potential causes for error covariance should be kept in mind. If the specification of correlated error terms within a model is supported by strong theoretical reasoning, then their inclusion would likely not be seen as an atheoretical attempt to improve model fit (MacCallum et al., 1992). Theory should be used to support the inclusion of correlated error terms to minimize the possibility that their inclusion is capitalizing on change.

While the results of the current study are confined to a very specific situation involving a single within-factor correlated error term, the levels of bias and the confidence interval attenuation warrant caution in more complex situations. Violations of the assumption of linear independence, wherever they appear in the model, should be treated with caution. The use of multiple item response formats alone presents the opportunity for error covariance not only within-factors, but also between-factors. As such, the increased complexity of assessments also increases the potential error covariance. With mounting research on method effects both in tests and surveys, researchers should exercise caution when evaluating the assumption of linear independence, until such time that research into more complex models has been conducted.

Limitations, Recommendations, and Future Research

There are several aspects of the current study that warrant further investigation. First, the results of a Monte Carlos study are limited to all of the specific conditions used for investigation (Hoogland & Boomsma, 1998). Though the 2-factor CFA model under consideration in this study is a common model within the measurement community, many aspects of and assumptions made in the study suggest the need for further investigation.

The various magnitudes specified for the factor loadings represented the likely range of factor loadings in applied research. At the same time, all factor loadings were set equal to each other within a given condition, an unlikely scenario given real data, but necessary to simplify the model to avoid a cumbersome number of total conditions. Error covariance between indicators with different loadings, and different loadings for indicators within the same factor as the effected items, may affect the levels of bias in a given indicator. Varying the magnitudes of the factor loadings would provide a clearer picture of what practitioners could expect for a given model if violations of linear dependence were possible.

The number of factors and indicators specified were meant to represent a likely scenario while maintaining simplicity for the sake of identifying factors contributing to or impeding bias. As structural equation modeling has become more common and widespread, the complexity of models, in terms of factors and indicators has increased. The increased complexity is readily seen in MTMM research (Marsh, 1990) and survey and test validation (Newman, Larsen, Cunningham, & Burkhart, 2015; Niehaus & Adelson, 2013). This increase in model complexity makes direct inference of the present results to these models difficult. As the complexity of the model increases, so do the factors potentially effecting presence of bias. Further research

should be undertaken to evaluate the impact of violations of linear dependence and the under-specification of correlated error terms on more complex and different types of models. Now that it is evident that violations of linear dependence can lead to biased estimates, it is important to directly determine how this might apply to more complex models and studies.

The potential combinations of sources of violations to linear independence present promising avenue of future research. The present study's simplifying condition of only correlating the error between one pair of items within the same factor limits the inferences that can be drawn about its effects on model validity. If a single pair of observed variables share some variance besides the specified latent factor, there is the possibility that more observed variables will also share error covariance. Further, the presence of error covariance need not be limited to within factor indicators. Items response formats, for instance, can be shared between observed variables loading on different latent factors, representing the potential for between-factor correlated error terms. The presence of bias caused by the failure to account for within-factor error covariance, combined with the failure of this bias to propagate to other factors, makes the presence of unaccounted-for between-factor error covariance warrants future investigation. The current research indicates that the failure to account for error covariance in a model biases loading estimates and coverage. At the same time, the location of the observed variables on different factors may limit the bias, since the bias does not appear to propagate across between-factor correlations.

Another situation, where a method effect, independent of the primary latent construct, causes error covariance across the majority of observed variables within a given model, may require a different approach than the one taken in the current study. While the specification of

correlated error terms may account for the error covariance, they may not provide the correct interpretation of factors affecting responses to the observed variables. In such cases, the specification of an additional factor to represent the method effect may be more appropriate. The use of a method factor in CFA models can be seen extensively in MTMM work by Marsh and Hocevar (1983) and others. Use of a separate "method" factor would allow for the direct evaluation and interpretation of the primary latent factor of interest, which may better represent the relationships among observed variables, and is worthy of further study on how it performs compared to correlated error models.

A related course of study, implied by the results of this study and other research, concerns the further and continued evaluation of method effects, especially item format. There is a general movement towards the inclusion of diverse item formats in assessments, as seen in the SBAS (Smarter Balanced Assessment Consortium, 2012). Just as the impact of error covariance among items warrants continued research, the factors inducing the error covariance demand attention. The presence of error covariance and violations of linear independence may go unnoticed if there is not strong understanding of the conditions causing it.

Knowledge of the factors yielding correlation between error terms can lead assessment makers to account for them in item specification, the rules used for creating items for the item pool. This in turn could lead to better items in the items pool with fewer rejected for issues surrounding their potential for error covariance. At the same time, assessment evaluators can use the knowledge of the factors likely to cause error covariance to aid in correctly specifying measurement models. The bias found in this study is avoidable, given good item writing practices, and knowledge of potential violations of linear dependence.

It is important to remember SEM models are almost always misspecified to some degree (MacCallum et al., 1992). Models are not expected to reproduce the real world exactly. A perfect understanding of all the possible relationships surrounding a latent construct is nearly impossible. The relationships between latent variables in path and structural models are determined by theory and research, as are the relationships between the indicators and latent variables. So too must the relationships between indicators be evaluated based on theory and research. There is an ever growing body of research on item formats, context and method effects. Assessment builders and users would do well to take this research into account, helping to avoid the potential pitfalls caused by error covariance.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, Calif.: Brooks/Cole Pub. Co.

American Psychological Association. (2015). Psychological Assessment. Retrieved 5/25/2015, from http://www.apa.org/pubs/journals/pas/

Anderson-Butcher, D., & Conroy, D. E. (2002). Factorial and Criterion Validity of Scores of a Measure of Belonging in Youth Development Programs. *Educational and Psychological Measurement, 62*(5), 857-876.

Bandalos, D. L., & Leite, W. (2013). Use of Monte Carlo Studies in Structural Equation Modeling Research. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (Vol. 1). Charlotte, NC: Information Age Publishing, Inc.

Bentler, P. M., & Yuan, K.-H. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate behavioral research, 34*(2), 181-197.

Bishop, G. F., Oldendick, R. W., & Tuchfarber, A. J. (1982). Political information processing: Question order and context effects. *Political Behavior, 4*(2), 177-200.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bollen, K. A. (2002). Latent Variables in Psychology and the Social Sciences. *Annual Review of Psychology, 53*(1).

Bostic, T. J., McGartland Rubio, D., & Hood, M. (2000). A Validation of the Subjective Vitality Scale Using Structural Equation Modeling. *Social Indicators Research Social Indicators Research : An International and Interdisciplinary Journal for Quality-of-Life Measurement, 52*(3), 313-324.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park: Sage Publications.

Bruning, R., Dempsey, M., Kauffman, D. F., McKim, C., & Zumbrunn, S. (2013). Examining dimensions of self-efficacy for writing. *Journal of Educational Psychology, 105*(1), 25-38. doi: 10.1037/a0029692

Byrne, B. M. (2006). *Structural equation modeling with EQS : basic concepts, applications, and programming*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Caldwell, D. J., & Pate, A. N. (2013). Effects of Question Formats on Student and Item Performance. *American Journal of Pharmaceutical Education, 77*(4), 1-5.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, N.J.: L. Erlbaum Associates.

Cole, D. A. (1987). Utility of confirmatory factor analysis in test validation research. *Journal of Consulting and Clinical Psychology, 55*(4), 584.

Cole, D. A., Ciesla, J. A., & Steiger, J. H. (2007). The insidious effects of failing to include design-driven correlated residuals in latent-variable covariance structure analysis. *Psychological Methods, 12*(4), 381-398. doi: 10.1037/1082-989X.12.4.381

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

DeVellis, R. F. (2012). *Scale development : theory and applications*. Thousand Oaks, Calif.: SAGE.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys : the tailored design method*. Hoboken, N.J.: Wiley & Sons.

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling, 12*(3), 343-367.

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate behavioral research, 42*(3), 509-529.

Fan, X., Wang, L., & Thompson, B. (1999). Effects of Sample Size, Estimation Methods, and Model Specification on Structural Equation Modeling Fit Indexes. *Structural Equation Modeling, 6*(1), 56-83.

Fleishman, J., & Benson, J. (1987). Using Lisrel to Evaluate Measurement Models and Scale Reliability. *Educational and Psychological Measurement, 47*(4), 925-939.

Gerbing, D. W., & Anderson, J. C. (1984). On the meaning of within-factor correlated measurement errors. *Journal of Consumer Research*.

Graf, E. A. (2009). Defining mathematics competency in the service of cognitively based assessment for grades 6 through 8: ETS Research Report No. RR-09-42). Princeton, NJ: ETS.

Green, S. B., & Hershberger, S. L. (2000). Correlated errors in true score models and their effect on coefficient alpha. *Structural Equation Modeling, 7*(2), 251-270.

Hallquist, M., & Wiley, J. (2014). MplusAutomation: Automating Mplus Model Estimation and Interpretation. R Package version 0.06-3. Retrieved from http://CRAN.R-project.org/package=MplusAutomation

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, Calif.: Sage Publications.

Hancock, G. R., & French, B. F. (2006). Power Analysis in Structural Equation Modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (Vol. 1): Iap.

Harvey, R. J., Billings, R. S., & Nilan, K. J. (1985). Confirmatory factor analysis of the Job Diagnostic Survey: Good news and bad news. *Journal of Applied Psychology, 70*(3), 461-468. doi: 10.1037/0021-9010.70.3.461

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101-125.

Heene, M., Hilbert, S., Freudenthaler, H. H., & Bühner, M. (2012). Sensitivity of SEM Fit Indexes With Respect to Violations of Uncorrelated Errors. *Structural Equation Modeling, 19*(1), 36-50. doi: 10.1080/10705511.2012.634710

Holyk, G. G. (2008). *United States leader and public support for multilateralism.* Available from http://worldcat.org /z-wcorg/ database.

Hoogland, J. J., & Boomsma, A. (1998). Robustness Studies in Covariance Structure Modeling. *Sociological Methods & Research, 26*(3).

Hoyle, R. H. (1995). *Structural equation modeling : concepts, issues, and applications*. Thousand Oaks: Sage Publications.

Hoyle, R. H., & Smith, G. T. (1994). Formulating clinical research hypotheses as structural equation models: a conceptual overview. *Journal of Consulting and Clinical Psychology, 62*(3), 429.

Hu, L.-t., & Bentler, P. M. (1995). Evaluating Model Fit. In R. H. Hoyle (Ed.), *Structural equation modeling : concepts, issues, and applications*. Thousand Oaks: Sage Publications.

Hu, L.-t., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424.

Hu, L.-t., & Bentler, P. M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus new alternatives. *Structural Equation Modeling, 6*(1).

IBM Corp. (2010). IBM SPSS Statistics for Windows (Version 19.0). Armonk, NY.

Kenney, S. R., Lac, A., Hummer, J. F., & LaBrie, J. W. (2014). Development and validation of the Hookup Motives Questionnaire (HMQ). *Psychological Assessment, 26*(4), 1127-1137. doi: 10.1037/a0037131

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological bulletin, 112*(1), 165-172. doi: 10.1037/0033-2909.112.1.165

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.

Kline, R. B. (2012). Assumptions in Structural Equation Modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling*. New York: Guilford Press.

Lee, S.-Y., & Bentler, P. (1980). Some Asymptotic Properties of Constrained Generalized Least-Squares Estimation in Covariance Sturcture Models. *South African Statistical Journal, 14*(2), 121-136.

Lewis, A. B., & Mayer, R. E. (1987). Students' miscomprehension of relational statements in arithmetic word problems. *Journal of Educational Psychology, 79*(4), 363-371. doi: 10.1037/0022-0663.79.4.363

Little, T. D. (2013). Longitudinal structural equation modeling.

Lomax, R. G. (2007). *An introduction to statistical concepts*. Mahwah, N.J.: Lawrence Erlbaum Associates Publishers.

MacCallum, R. C. (2003). Working with Imperfect Models. *Multivariate behavioral research, 38*(1), 113-139. doi: 10.1207/S15327906MBR3801_5

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological bulletin, 111*(3), 490-504. doi: 10.1037/0033-2909.111.3.490

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological bulletin, 114*(1), 185-199. doi: 10.1037/0033-2909.114.1.185

Mack, J., Herrberg, M., Hetzel, A., Wallesch, C. W., Bengel, J., Schulz, M., . . . Schönberger, M. (2015). The factorial and discriminant validity of the German version of the Post-traumatic Growth Inventory in stroke patients. *Neuropsychological rehabilitation, 25*(2), 216-232.

Marsh, H. W. (1990). Confirmatory Factor Analysis of Multitrait-Multimethod Data: The Construct Validation of Multidimensional Self-Concept Responses. *Journal of Personality, 58*(4), 661-692. doi: 10.1111/1467-6494.ep9103184273

Marsh, H. W., Abduljabbar, A. S., Abu-Hilal, M. M., Morin, A. J. S., Abdelfattah, F., Leung, K. C., . . . Parker, P. (2013). Factorial, convergent, and discriminant validity of timss math and science motivation measures: A comparison of Arab and Anglo-Saxon countries. *Journal of Educational Psychology, 105*(1), 108-128. doi: 10.1037/a0029907

Marsh, H. W., & Bailey, M. (1991). Confirmatory Factor Analyses of Multitrait-Multimethod Data: A Comparison of Alternative Models. *Applied Psychological Measurement Applied Psychological Measurement, 15*(1), 47-70.

Marsh, H. W., Byrne, B. M., & Craven, R. (1992). Overcoming problems in confirmatory factor analyses of MTMM data: The correlated uniqueness model and factorial invariance. *Multivariate behavioral research, 27*(4), 489-507.

Marsh, H. W., & Hocevar, D. (1983). Confirmatory Factor Analysis of Multitrait-Multimethod Matrices. *Journal of Educational Measurement, 20*(3), 231-248. doi: 10.2307/1434714

Marsh, H. W., & Hocevar, D. (1988). A new, more powerful approach to multitrait-multimethod analyses: Application of second-order confirmatory factor analysis. *Journal of Applied Psychology, 73*(1), 107-117. doi: 10.1037/0021-9010.73.1.107

Matos, M., Pinto-Gouveia, J., Gilbert, P., Duarte, C., & Figueiredo, C. (2015). The Other As Shamer Scale – 2: Development and validation of a short version of a measure of external shame. *Personality and Individual Differences, 74*(0), 6-11. doi: http://dx.doi.org/10.1016/j.paid.2014.09.037

Maul, A. (2013). Method Effects and the Meaning of Measurement. *Frontiers in Psychology, 4*, 169. doi: 10.3389/fpsyg.2013.00169

Messick, S. (1990). *Validity of test interpretation and use*. Princeton, N.J.: Educational Testing Service.

Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 2*(3), 255-273.

Mooney, C. Z. (1997). *Monte Carlo simulation*. Thousand Oaks, Calif.: Sage Publications.

Muthén, B. O., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52*(3), 431-462.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus User's Guide. Seventh Edition.* Los Angeles, Ca: Muthén & Muthén.

Newman, J. L. E., Larsen, J. L., Cunningham, K. B., & Burkhart, B. R. (2015). An Examination of the Factor Structure of the Millon Adolescent Clinical Inventory in a Sample of Detained Adolescent Boys. *Psychological Assessment*. doi: 10.1037/a0038779.supp (Supplemental)

Niehaus, K., & Adelson, J. L. (2013). Self-concept and native language background: A study of measurement invariance and cross-group comparisons in third grade. *Journal of Educational Psychology, 105*(1), 226-240. doi: 10.1037/a0030556

Norwalk, K. E., DiPerna, J. C., & Lei, P.-W. (2014). Confirmatory factor analysis of the Early Arithmetic, Reading, and Learning Indicators (EARLI). *Journal of School Psychology, 52*(1), 83-96. doi: http://dx.doi.org/10.1016/j.jsp.2013.11.006

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo Experiments: Design and Implementation. *Structural Equation Modeling, 8*(2), 287-312.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879-903. doi: 10.1037/0021-9010.88.5.879

Preacher, K. J., & Coffman, D. L. (2006, May). Computing power and minumum sample size for RMSEA. Retrieved from http://quantpsy.org

Purpura, D. J., & Lonigan, C. J. (2013). Informal Numeracy Skills: The Structure and Relations Among Numbering, Relations, and Arithmetic Operations in Preschool. *American Educational Research Journal, 50*(1), 178-209. doi: 10.3102/0002831212465332

R Core Team. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org

Raykov, T. (2012). Scale Construction and Development Using Structural Equation Modeling. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling*. New York: Guilford Press.

Richards, K. A. R., Levesque-Bristol, C., & Templin, T. J. (2014). Initial Validation of the Teacher/Coach Role Conflict Scale. *Measurement in Physical Education & Exercise Science, 18*(4), 259-272. doi: 10.1080/1091367X.2014.932283

Robert, C., & Casella, G. (2013). *Monte Carlo statistical methods*: Springer Science & Business Media.

Rozeboom, W. W. (1989). The reliability of a linear composite of nonequivalent subtests. *Applied Psychological Measurement, 13*(3), 277-283.

Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling, 16*(4), 561-582.

Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys : experiments on question form, wording, and context*. Thousand Oaks, CA: Sage Publications.

Smarter Balanced Assessment Consortium. (2012). *Prelimary Test Blueprints*. Retrieved from Smarter Balanced Assessment Consortium website: http://www.smarterbalanced.org/wordpress/wp-content/uploads/2011/12/Smarter-Balanced-Preliminary-Test-Blueprints.pdf.

Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology, 15*(2), 201-292. doi: 10.2307/1412107

Stevens, J. (2009). *Applied multivariate statistics for the social sciences*. New York: Routledge.

Taylor, C. S. (2013). *Validity and validation*: Oxford University Press.

Taylor, C. S., & Lee, Y. (2011). Ethnic DIF in Reading Tests With Mixed Item Formats. *Educational Assessment, 16*(1), 35-68.

Taylor, C. S., & Lee, Y. (2012). Gender DIF in Reading and Mathematics Tests With Mixed Item Formats. *Applied Measurement in Education, 25*(3), 246-280. doi: 10.1080/08957347.2012.687650

Taylor, C. S., & Nolen, S. B. (2005). *Classroom assessment : supporting teaching and learning in real classrooms*. Upper Saddle River, N.J.: Pearson/Merrill/Prentice Hall.

Thompson-Schill, S. L., Kurtz, K. J., & Gabrieli, J. D. E. (1998). Effects of Semantic and Associative Relatedness on Automatic Priming. *Journal of Memory and Language, 38*(4), 440-458.

Thurber, R. S., Shinn, M. R., & Smolkowski, K. (2002). What is measured in mathematics tests? Construct validity of curriculum-based mathematics measures. *School Psychology Review, 31*(4), 498-513.

Todorov, A. (2000). Context Effects in National Health Survey. *Public Opinion Quarterly, 64*(1), 65-76.

Tourangeau, R., Singer, E., & Presser, S. (2003). Context Effects in Attitude Surveys: Effects on Remote Items and Impact on Predictive Validity. *Sociological Methods & Research, 31*(4), 486-513.

Uhan, S., & Fink, M. H. (2013). Context effects in social surveys: between instrument and respondent. *Teor. Praksa Teorija in Praksa, 50*(1), 233-248.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software, 21*(12), 1-20.

Wilson, D., Jones, D., Bocell, F., Kim, M. J., Veilleux, N., Floyd-Smith, T., . . . Plett, M. (In Press). Belonging and Engagement among Undergraduate STEM Students: A Multi-institutional Study. *Research in Higher Education*.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 73*(6), 913-934.

Zimmerman, D. W., Zumbo, B. D., & Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educational and Psychological Measurement, 53*(1), 33-49.

Table 1. Overall Descriptive Statistics across Conditions for Selected Indicators

Indicator		Model Estimate	*Mean*	*(SD)*	*Mean*	*(SD)*	Model Estimate		Indicator
	X1	Loading Estimate	0.711	(0.122)	0.711	(0.122)	Loading Estimate	X2	
		95% CI Coverage	0.691	(0.299)	0.691	(0.299)	95% CI Coverage		
		Raw Bias	0.111	(0.102)	0.111	(0.102)	Raw Bias		
		Relative Bias	0.221	(0.243)	0.221	(0.243)	Relative Bias		
	X3	Loading Estimate	0.535	(0.171)	0.535	(0.171)	Loading Estimate	X4	
		95% CI Coverage	0.807	(0.186)	0.807	(0.186)	95% CI Coverage		
		Raw Bias	-0.065	(0.053)	-0.065	(0.053)	Raw Bias		
		Relative Bias	-0.124	(0.122)	-0.124	(0.122)	Relative Bias		
	X5	Loading Estimate	0.599	(0.138)	0.599	(0.138)	Loading Estimate	X6	
		95% CI Coverage	0.944	(0.005)	0.944	(0.005)	95% CI Coverage		
		Raw Bias	-0.001	(0.005)	-0.001	(0.005)	Raw Bias		
		Relative Bias	0.000	(0.012)	0.000	(0.012)	Relative Bias		

Note. 36,0000,000 replicates for each of the collapsed conditions shown above. X1 is indicator with violation to linear independence.

Table 2. Descriptive Statistics by Sample Size for Selected Indicators

	Sample Size		80		160		400	
	Model Estimate		*Mean*	*(SD)*	*Mean*	*(SD)*	*Mean*	*(SD)*
Indicator	X1	Loading Estimate	0.711	(0.122)	0.711	(0.123)	0.711	(0.124)
		95% CI Coverage	0.691	(0.299)	0.714	(0.261)	0.533	(0.369)
		Raw Bias	0.111	(0.102)	0.111	(0.101)	0.111	(0.099)
		Relative Bias	0.221	(0.243)	0.220	(0.242)	0.218	(0.234)
	X3	Loading Estimate	0.535	(0.171)	0.535	(0.172)	0.536	(0.173)
		95% CI Coverage	0.807	(0.186)	0.831	(0.131)	0.711	(0.261)
		Raw Bias	-0.065	(0.053)	-0.065	(0.054)	-0.064	(0.054)
		Relative Bias	-0.124	(0.122)	-0.125	(0.124)	-0.124	(0.123)
	X5	Loading Estimate	0.599	(0.138)	0.599	(0.140)	0.599	(0.141)
		95% CI Coverage	0.944	(0.005)	0.944	(0.002)	0.948	(0.001)
		Raw Bias	-0.001	(0.005)	-0.001	(0.003)	-0.001	(0.001)
		Relative Bias	0.000	(0.012)	-0.002	(0.006)	-0.001	(0.001)

Note. 12,000,000 replicates for each of the collapsed conditions shown above.
X1 is indicator with violation to linear independence.

Table 3. Descriptive Statistics by Loading Magnitude for Selected Indicators

Factor Loading Mag			0.40		0.50		0.60		0.70		0.80	
Model Estimate			*Mean*	*(SD)*	*Mean*	*(SD)*	*Mean*	*(SD)*	*Mean*	*(SD)*	*Mean*	*(SD)*
Indicator	X1	Loading Estimate	0.598	(0.131)	0.643	(0.101)	0.702	(0.076)	0.769	(0.053)	0.841	(0.034)
		95% CI Coverage	0.617	(0.326)	0.606	(0.335)	0.652	(0.315)	0.739	(0.262)	0.843	(0.164)
		Raw Bias	0.198	(0.131)	0.143	(0.101)	0.102	(0.076)	0.069	(0.053)	0.041	(0.034)
		Relative Bias	0.496	(0.328)	0.285	(0.202)	0.171	(0.126)	0.099	(0.076)	0.052	(0.042)
	X3	Loading Estimate	0.306	(0.066)	0.417	(0.059)	0.534	(0.048)	0.652	(0.035)	0.767	(0.023)
		95% CI Coverage	0.734	(0.234)	0.759	(0.219)	0.801	(0.184)	0.849	(0.131)	0.891	(0.073)
		Raw Bias	-0.094	(0.066)	-0.083	(0.059)	-0.066	(0.048)	-0.048	(0.035)	-0.033	(0.023)
		Relative Bias	-0.235	(0.166)	-0.165	(0.117)	-0.110	(0.080)	-0.069	(0.051)	-0.041	(0.029)
	X5	Loading Estimate	0.407	(0.008)	0.499	(0.001)	0.597	(0.002)	0.696	(0.002)	0.796	(0.002)
		95% CI Coverage	0.940	(0.006)	0.942	(0.005)	0.944	(0.004)	0.946	(0.002)	0.947	(0.002)
		Raw Bias	0.007	(0.008)	-0.001	(0.001)	-0.003	(0.002)	-0.004	(0.002)	-0.004	(0.002)
		Relative Bias	0.016	(0.019)	-0.002	(0.002)	-0.005	(0.003)	-0.005	(0.003)	-0.005	(0.003)

Note. 7,200,000 replicates for each of the collapsed conditions shown above. X1 is indicator with violation to linear independence.

Table 4. Descriptive Statistics of Model Outcomes by Between-Factor Correlation Magnitude for Selected Indicators

Factor Correlation Mag			0.00		0.10		0.30		0.50	
Model Estimate			*Mean*	*(SD)*	*Mean*	*(SD)*	*Mean*	*(SD)*	*Mean*	*(SD)*
Indicator	X1	Loading Estimate	0.714	(0.122)	0.713	(0.122)	0.711	(0.123)	0.706	(0.124)
		95% CI Coverage	0.693	(0.298)	0.693	(0.298)	0.690	(0.301)	0.689	(0.305)
		Raw Bias	0.114	(0.104)	0.113	(0.104)	0.111	(0.102)	0.106	(0.100)
		Relative Bias	0.226	(0.248)	0.226	(0.248)	0.220	(0.244)	0.210	(0.237)
	X3	Loading Estimate	0.532	(0.173)	0.533	(0.173)	0.535	(0.171)	0.540	(0.169)
		95% CI Coverage	0.799	(0.194)	0.800	(0.193)	0.807	(0.187)	0.821	(0.174)
		Raw Bias	-0.068	(0.055)	-0.067	(0.055)	-0.065	(0.053)	-0.060	(0.050)
		Relative Bias	-0.129	(0.126)	-0.129	(0.126)	-0.124	(0.122)	-0.114	(0.115)
	X5	Loading Estimate	0.600	(0.138)	0.599	(0.138)	0.599	(0.139)	0.598	(0.140)
		95% CI Coverage	0.944	(0.004)	0.944	(0.004)	0.944	(0.005)	0.943	(0.005)
		Raw Bias	0.000	(0.006)	-0.001	(0.006)	-0.001	(0.005)	-0.002	(0.004)
		Relative Bias	0.001	(0.015)	0.001	(0.014)	-0.001	(0.011)	-0.002	(0.008)

Note. 9,000,000 replicates for each of the collapsed conditions shown above. X1 is indicator with violation to linear independence.

Table 5. Descriptive Statistics of Model Outcomes by Error Correlation Magnitude for Selected Indicators

Error Correlation Mag			0.00		0.10		0.20		0.30		0.40		0.50	
Model Estimate			*Mean*	*(SD)*	*Mean*	*(SD)*	*Mean*	*(SD)*	*Mean*	*(SD)*	*Mean*	*(SD)*	*Mean*	*(SD)*
Indicator	X1	Loading Estimate	0.599	(0.139)	0.640	(0.119)	0.686	(0.096)	0.735	(0.075)	0.781	(0.057)	0.825	(0.044)
		95% CI Coverage	0.944	(0.005)	0.923	(0.025)	0.820	(0.129)	0.649	(0.246)	0.479	(0.284)	0.333	(0.268)
		Raw Bias	-0.001	(0.005)	0.040	(0.025)	0.086	(0.049)	0.135	(0.070)	0.181	(0.088)	0.225	(0.101)
		Relative Bias	-0.001	(0.011)	0.081	(0.068)	0.174	(0.136)	0.269	(0.200)	0.359	(0.254)	0.442	(0.298)
	X3	Loading Estimate	0.599	(0.139)	0.577	(0.150)	0.551	(0.163)	0.522	(0.174)	0.494	(0.181)	0.468	(0.183)
		95% CI Coverage	0.944	(0.005)	0.930	(0.011)	0.887	(0.049)	0.805	(0.118)	0.696	(0.186)	0.578	(0.231)
		Raw Bias	-0.001	(0.005)	-0.023	(0.009)	-0.049	(0.021)	-0.078	(0.032)	-0.106	(0.038)	-0.132	(0.041)
		Relative Bias	-0.001	(0.012)	-0.043	(0.027)	-0.096	(0.063)	-0.151	(0.095)	-0.203	(0.117)	-0.250	(0.130)
	X5	Loading Estimate	0.599	(0.139)	0.599	(0.139)	0.599	(0.139)	0.599	(0.139)	0.599	(0.139)	0.599	(0.139)
		95% CI Coverage	0.943	(0.005)	0.944	(0.005)	0.944	(0.005)	0.944	(0.005)	0.944	(0.004)	0.944	(0.004)
		Raw Bias	-0.001	(0.005)	-0.001	(0.005)	-0.001	(0.006)	-0.001	(0.006)	-0.001	(0.006)	-0.001	(0.006)
		Relative Bias	0.000	(0.012)	0.000	(0.012)	0.000	(0.012)	0.000	(0.012)	0.000	(0.012)	0.000	(0.013)

Note. 6,000,000 replicates per condition shown above. X1 is indicator with violation to linear independence.

Table 6. Results for 4-Factor ANOVAS with All 2-Way Interactions on Relative Bias in Selected Indicators

	Relative Bias in X1			Relative Bias in X3			Relative Bias in X5		
Source	*F*	*p*	ω^2	*F*	*p*	ω^2	*F*	*p*	ω^2
Intercept	456817.560	< 0.001	--	259203.145	< 0.001	--	2.068	> 0.05	--
Sample Size	28.661	< 0.001	< 0.001	7.294	< 0.001	< 0.001	99.978	< 0.001	**0.019**
Loading	59058.446	< 0.001	**0.426**	20278.285	< 0.001	**0.324**	1219.635	< 0.001	**0.477**
Factor Corr.	128.319	< 0.001	0.001	213.124	< 0.001	0.003	42.441	< 0.001	**0.012**
Error Corr.	44124.805	< 0.001	**0.398**	25506.391	< 0.001	**0.509**	0.608	> 0.05	< 0.001
Sample Size X Loading	119.549	< 0.001	0.002	62.019	< 0.001	0.002	540.036	< 0.001	**0.422**
Sample Size X Factor Corr.	6.107	< 0.001	< 0.001	0.177	> 0.05	< 0.001	14.985	< 0.001	0.008
Sample Size X Error Corr.	4.426	< 0.001	< 0.001	2.803	< 0.01	< 0.001	0.266	> 0.05	< 0.001
Loading X Factor Corr.	25.109	< 0.001	0.001	17.163	< 0.001	0.001	26.422	< 0.001	**0.031**
Loading X Error Corr.	4759.871	< 0.001	**0.172**	1992.907	< 0.001	**0.159**	0.475	> 0.05	0.001
Factor Corr. X Error Corr.	4.481	< 0.001	< 0.001	16.813	< 0.001	0.001	0.280	> 0.05	< 0.001

Note. 100,000 replicates per condition, with 360 conditions. Significant effects with $p < .0083$ and substantive effect sizes in boldface. X1 is indicator with violation to linear independence.

Table 7. Results for 4-Factor ANOVA with all 2-Way Interactions on 95% CI Coverage in Selected Indicators

	95% CI Coverage for X1			95% CI Coverage for X3			95% CI Coverage for X5		
Source	*F*	*p*	ω^2	*F*	*p*	ω^2	*F*	*p*	ω^2
Intercept	93333.51	< 0.001	--	251646.9	< 0.001	--	492413659.2	< 0.001	--
Sample Size	1420.666	< 0.001	**0.163**	961.262	< 0.001	**0.143**	3441.517	< 0.001	**0.594**
Loading	386.022	< 0.001	**0.088**	317.839	< 0.001	**0.095**	782.108	< 0.001	**0.270**
Factor Corr.	0.221	> 0.05	< 0.001	10.178	< 0.001	0.002	24.570	< 0.001	0.006
Error Corr.	2014.622	< 0.001	**0.577**	1355.085	< 0.001	**0.506**	2.710	< 0.05	< 0.001
Sample Size X Loading	17.457	< 0.001	0.008	43.270	< 0.001	**0.025**	118.508	< 0.001	**0.082**
Sample Size X Factor Corr.	0.372	> 0.05	< 0.001	1.692	> 0.05	< 0.001	8.638	< 0.001	0.004
Sample Size X Error Corr.	167.986	< 0.001	**0.096**	188.404	< 0.001	**0.140**	2.487	< 0.05	0.002
Loading X Factor Corr.	1.147	> 0.05	0.001	0.350	> 0.05	< 0.001	12.299	< 0.001	**0.012**
Loading X Error Corr.	43.663	< 0.001	**0.050**	42.623	< 0.001	**0.063**	1.549	> 0.05	0.002
Factor Corr. X Error Corr.	0.118	> 0.05	< 0.001	1.041	> 0.05	0.001	0.403	> 0.05	< 0.001

Note. 100,000 replicates per condition, with 360 conditions. Significant effects with $p < .0083$ and substantive effect sizes in boldface. X1 is indicator with violation to linear independence.

Table 8. Model Fit Statistics for Applied Analysis

Fit Statistic	No Correlated Error Term	Correlated Error Term
χ^2(df)	90.189*(19)	60.648*(18)
RMSEA	0.065	0.052
SRMR	0.024	0.018
CFI	0.980	0.988
TLI	0.970	0.981

*$p < 0.001$.

Table 9. Applied Analysis Results

Variable	Correlated Error			No Correlated Error			Bias	
	Loading	*Std. Loading*	*(SE)*	*Loading*	*Std. Loading*	*(SE)*	*Raw*	*Relative*
F1: Major Belonging								
B09*	1.000	0.788	(0.019)	1.000	0.841	(0.015)	0.053	0.067
A01*	0.916	0.664	(0.021)	0.945	0.731	(0.021)	0.067	0.101
A20	0.912	0.714	(0.025)	0.823	0.687	(0.019)	-0.027	-0.038
A16	1.017	0.767	(0.019)	0.914	0.736	(0.019)	-0.031	-0.040
F2: Class Belonging								
I15	1.000	0.810	(0.015)	1.000	0.811	(0.015)	0.001	0.001
I20	0.959	0.781	(0.016)	0.958	0.780	(0.016)	-0.001	-0.001
I18	0.984	0.845	(0.013)	0.985	0.846	(0.013)	0.001	0.001
I09	0.937	0.772	(0.016)	0.936	0.771	(0.016)	-0.001	-0.001
Correlations								
Class with Major	0.338	0.657	(0.026)	0.351	0.640	(0.026)	-0.017	-0.026
A01 with B09	0.112	0.279	(0.043)	---	---	---	---	---

Note. Asterisks indicate the two items with correlated error. Bias Calculated for Standardized Loadings; Std. = Standardized; SE are for standardized loadings; all estimates significant at $p < 0.001$.

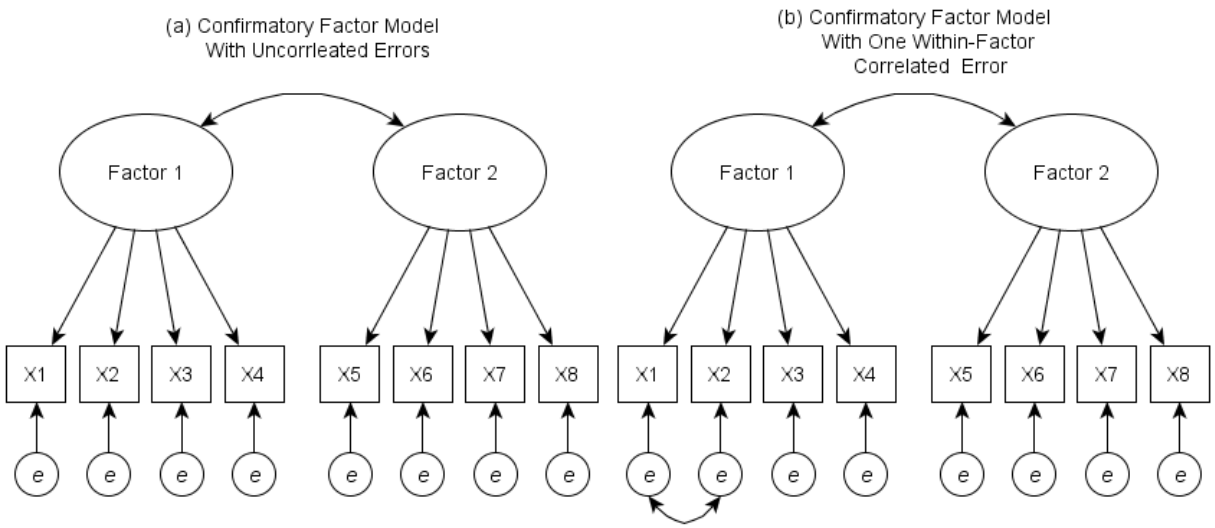

Figure 1: Examples of Confirmatory Factor Analysis (CFA) models

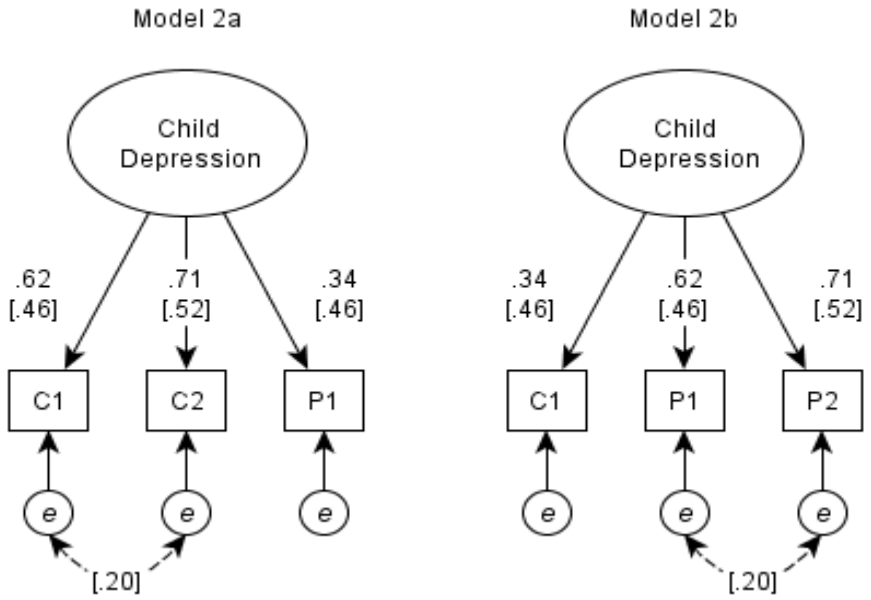

Figure 2. Example of conceptually misspecified model adapted from Cole et al. (2007)

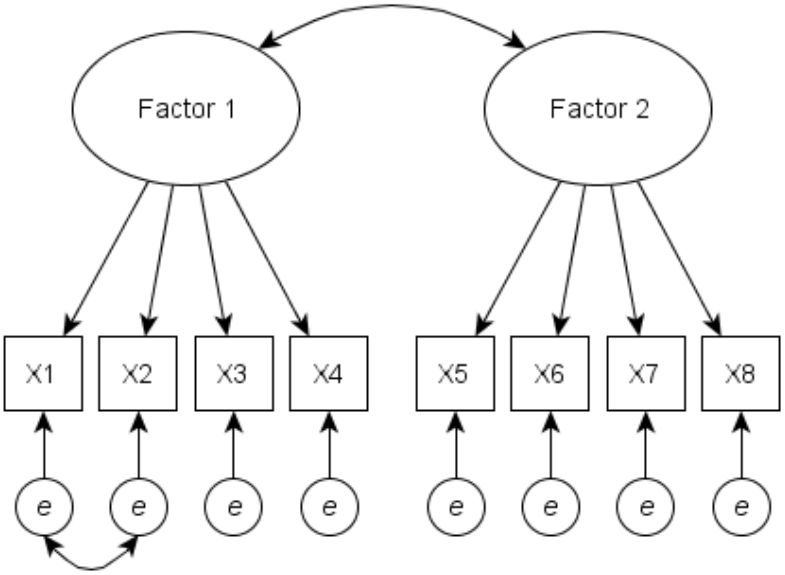

Figure 3. Example of a 2-factor CFA model with four indicators per factor and one correlated error term

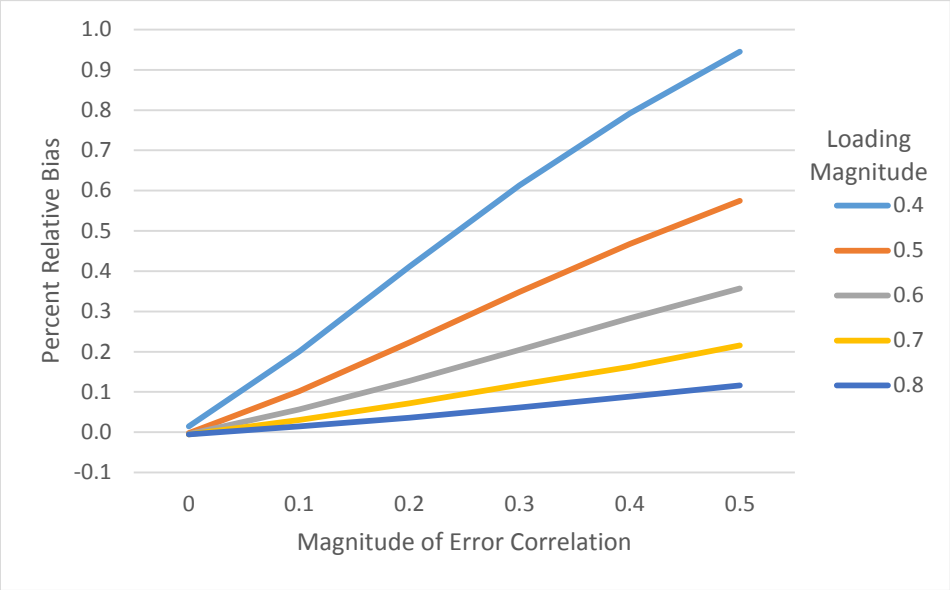

Figure 4. Magnitude of Mean Relative Bias In X1 by Correlated Error and Loading Magnitude

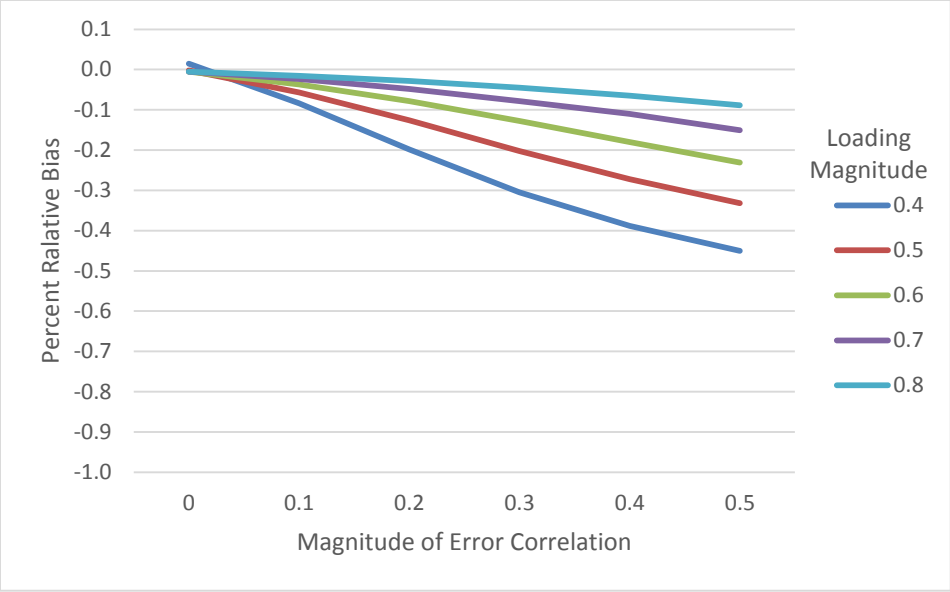

Figure 5. Magnitude of Mean Relative Bias in X3 by Correlated Error and Loading Magnitude

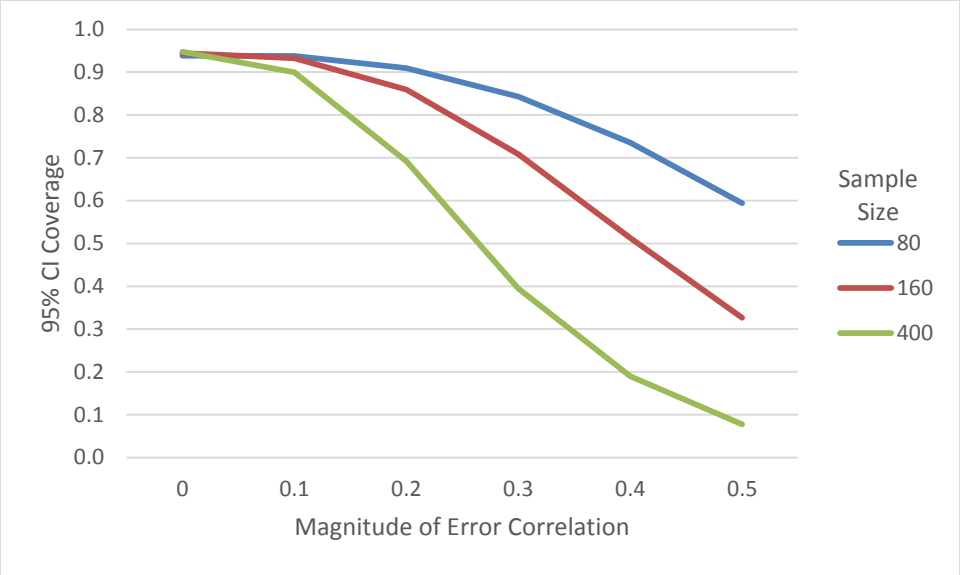

Figure 6. Mean 95% CI Coverage for X1 by Correlated Error and Sample Size

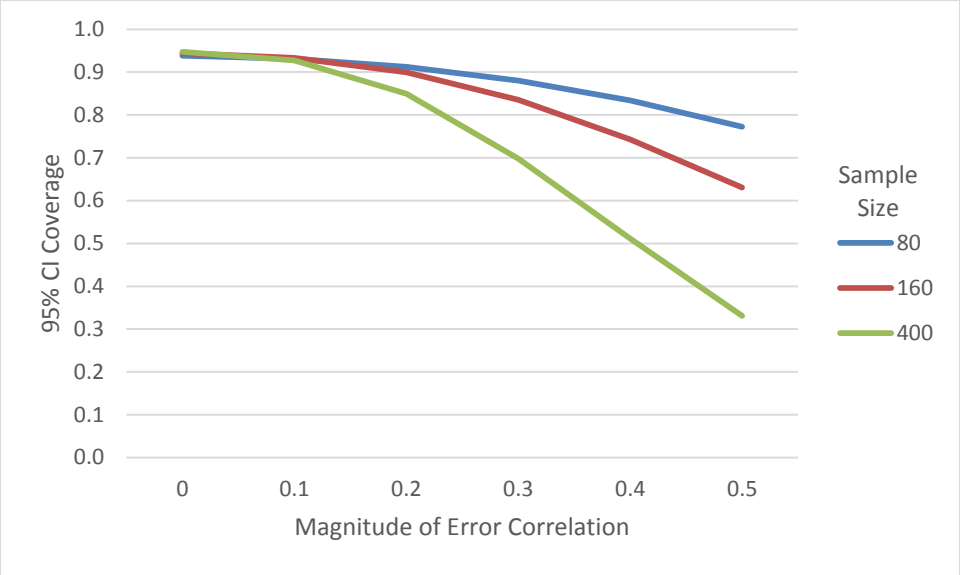

Figure 7. Mean 95% CI Coverage For X3 by Correlated Error and Sample Size

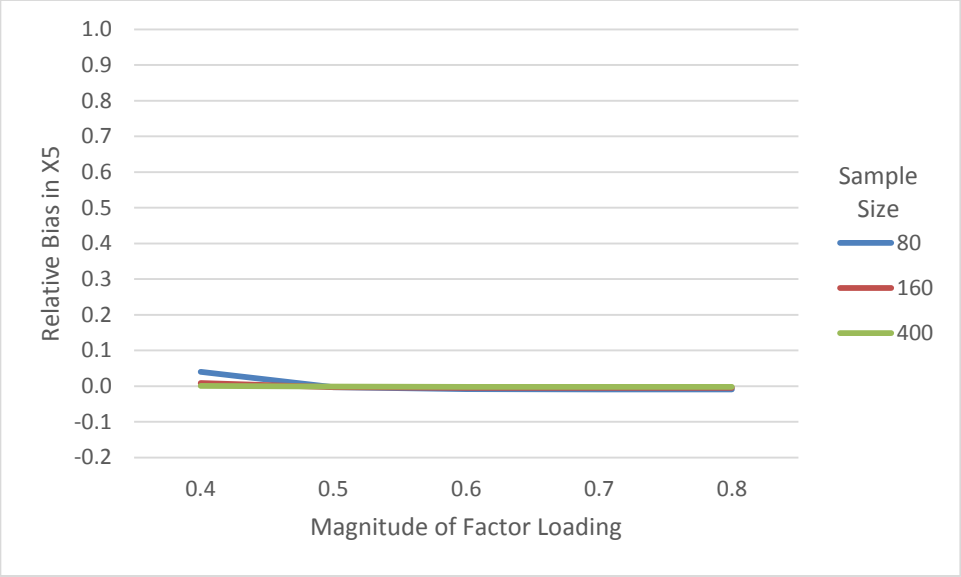

Figure 8. Mean Relative Bias in X5 by Loading Magnitude and Sample Size

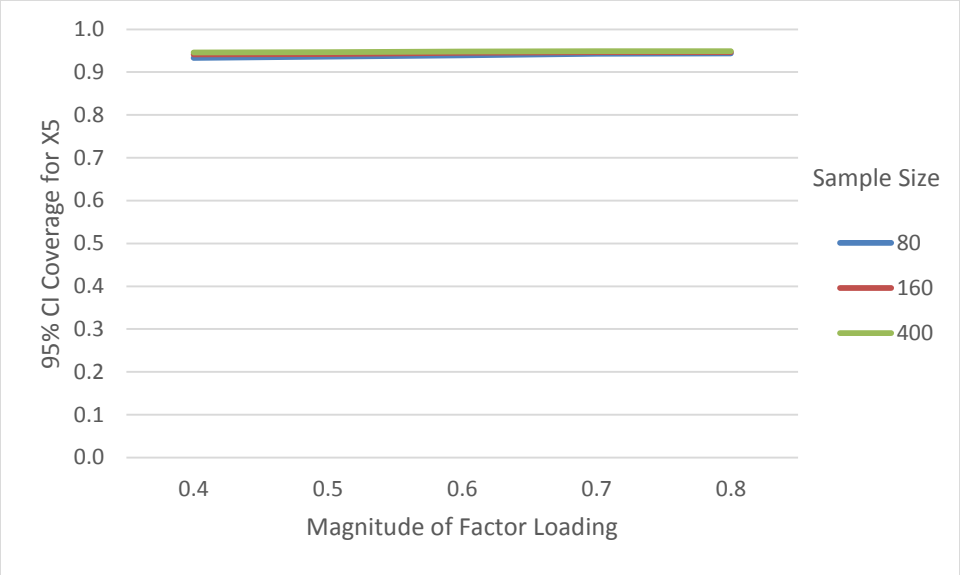

Figure 9. Mean 95% CI Coverage for X5 by Loading Magnitude and Sample Size

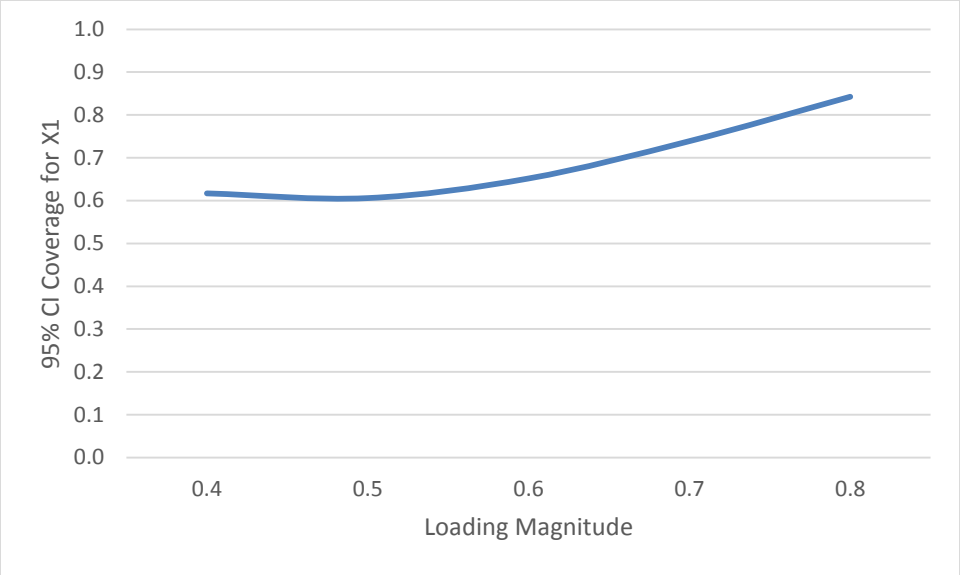

Figure 10. Mean 95% CI Coverage for X1 by loading magnitude

Appendix I. Example of Single *Mplus* Data Simulation (Input file)

```
Title: Monte Carlo w/ .4 loadings, .0 factor corr, .0 error corr, n=80
MONTECARLO:
  NAMES ARE x1-x8;
  NOBSERVATIONS = 80;
  NREPS = 100000;
  SEED = 80040;
  RESULTS = Rep04080.dat;
MODEL POPULATION:
  F1 BY x1-x4*.4;
  F2 BY x5-x8*.4;
  F1@1 F2@1;
  x1-x8*.84;
  F1 WITH F2*.0;
  x1 WITH x2*.0;
MODEL:
  F1 BY x1-x4*.4;
  F2 BY x5-x8*.4;
  F1-F2@1;
  x1-x8*.84;
  F1 WITH F2*.0;
ANALYSIS: ESTIMATOR = ML;
OUTPUT: TECH9;
```

Appendix II. Example of Input File for "Automated Mplus" in R

```
[[init]]
iterators = fcorr load n;
fcorr = 0 1 3 5;
load = 4 5 6 7 8;
n = 080 160 400;
res#load = .84 .75 .64 .51 .36;
cu#load = 0 0 0 0 0; !must be run once for each error correlation level
filename = "rep[[fcorr]][[load]]0[[n]].inp";
outputDirectory = "C:/FraserBocell/Dissertation/Input";
[[/init]]

Title: Testing file generation
MONTECARLO:
  NAMES ARE x1-x8;
  NOBSERVATIONS = [[n]];
  NREPS = 100000;
  SEED = [[n]][[fcorr]][[load]][[cu#load]];
  RESULTS = Rep[[fcorr]][[load]][[cu#load]][[n]].dat;
MODEL POPULATION:
  F1 BY x1-x4*.[[load]];
  F2 BY x5-x8*.[[load]];
  F1@1 F2@1;
  x1-x8*[[res#load]];
  F1 WITH F2*.[[fcorr]];
  x1 WITH x2*.[[cu#load]];
MODEL:
  F1 BY x1-x4*.[[load]];
  F2 BY x5-x8*.[[load]];
  F1-F2@1;
  x1-x8*[[res#load]];
  F1 WITH F2*.[[fcorr]];
ANALYSIS: ESTIMATOR = ML;
OUTPUT: TECH9;
```

Appendix III. *R* Code for Results Saving from "Automated Mplus"

```
#Load MplusAutomation
library(MplusAutomation)

#Create Model inputs
createModels("C:/FraserBocell/Dissertation/InpGen0.txt")
createModels("C:/FraserBocell/Dissertation/InpGen1.txt")
createModels("C:/FraserBocell/Dissertation/InpGen2.txt")
createModels("C:/FraserBocell/Dissertation/InpGen3.txt")
createModels("C:/FraserBocell/Dissertation/InpGen4.txt")
createModels("C:/FraserBocell/Dissertation/InpGen5.txt")

#Run model inputs from a specified folder
runModels("C:/FraserBocell/Dissertation/Input/80")
runModels("C:/FraserBocell/Dissertation/Input/160")
runModels("C:/FraserBocell/Dissertation/Input/400")

#extract parameters from .out files in specified folder
#run through all syntax for each sample size seperately
allmodelparameters <- extractModelParameters("C:/FraserBocell/Dissertation/Input/80")
allmodelparameters <- extractModelParameters("C:/FraserBocell/Dissertation/Input/160")
allmodelparameters <- extractModelParameters("C:/FraserBocell/Dissertation/Input/400")

summary(allmodelparameters)

#all of the following is run once per sample size, 80, 160, 400
#retain only unstandardized results
standardizedOnly <- sapply(allmodelparameters, "[", "standardized")

#retain existing file names without appending .standardized
oldnames <- names(allmodelparameters)
standardizedOnly <- sapply(allmodelparameters, "[", "standardized")
names(standardizedOnly) <- oldnames


#Add the filename as a field in the data.frame
lapply(names(standardizedOnly), function(element) {standardizedOnly[[element]]$filename <<-
element})

#this will only work if all data.frames have identical columns
combinedparameters <- do.call("rbind", standardizedOnly)
```

```
#get rid of row names
rownames(combinedparameters) = NULL

#combine paramHeader and Param
combinedparameters$parameter <- paste0(combinedparameters$paramHeader,
combinedparameters$param)

#Delete paramHeader and Param
combinedparameters$paramHeader = NULL
combinedparameters$param = NULL

#melt data
library(reshape2)
results <- melt(combinedparameters, id = c("filename", "parameter"), measured =
c("population", "average", "population_sd", "average_se", "mse", "cover_95", "pct_sig_coef"))

#combine parameter and variable
results$variables <- paste0(results$parameter, results$variable)

#delete parameter and variables
results$variable = NULL
results$parameter = NULL

#cast data, run one at a time
final80 <- dcast(results, filename ~ variables)
fina160 <- dcast(results, filename ~ variables)
fina400 <- dcast(results, filename ~ variables)

#export to excel
write.csv(final80, "C:/FraserBocell/Dissertation/averageresults80.csv")
write.csv(final160, "C:/FraserBocell/Dissertation/averageresults160.csv")
write.csv(final400, "C:/FraserBocell/Dissertation/averageresults400.csv")

citation(package = "MplusAutomation")
citation(package = "reshape2")
```

Appendix IV. Example of Single *Mplus* Applied Analysis (Input file)

```
Title: Belonging Class and Belonging Major with CU
DATA: FILE IS H:\Dissertation\UW noFresh full model all indicators.dat;
Variable: NAMES ARE A01
A16
A20
B09
I09
I15
I18
I20;
MISSING ARE  A01
A16
A20
B09
I09
I15
I18
I20 (99);
Model: BClss By I15 I20 I18 I09;
BMjr By B09 A20 A01 A16;
A01 WITH A20;
OUTPUT: TECH4;
MODINDICES;
STANDARDIZED (STDYX);
```