7-2013

Structural Equation Models with Small Samples: A Comparative Study of Four Approaches

Frances L. Chumney
University of Nebraska-Lincoln, franchumney@hotmail.com

STRUCTURAL EQUATION MODELS WITH SMALL SAMPLES: A COMPARATIVE STUDY OF FOUR APPROACHES

by

Frances L. Chumney

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Psychological Studies in Education

Under the Supervision of Professor James A. Bovaird

Lincoln, Nebraska

July 15, 2013

STRUCTURAL EQUATION MODELS WITH SMALL SAMPLES: A COMPARATIVE STUDY OF FOUR APPROACHES

Frances L. Chumney, Ph.D.

University of Nebraska, 2013

Adviser: James A. Bovaird

The purpose of this study was to evaluate the performance of estimation methods (Maximum Likelihood, Partial Least Squares, Generalized Structured Components Analysis, Markov Chain Monte Carlo) when applied to structural equation models with small samples. Trends in educational and social science research require scientists to investigate increasingly complex phenomena with regard for the contextual factors which influence their occurrence and change. These additional layers of exploration lead to complex hypotheses and require advanced analytic approaches such as structural equation modeling. A mismatch exists between analytic technique and the realities of applied research. Structural equation modeling requires large samples in general and even larger samples for complex models; for applied researchers, large samples are often difficult and even impossible to obtain. The unique contribution of this study is the simultaneous evaluation of these four estimation methods to determine the analytic conditions under which each method might be of value to researchers. A simulation study with a 3×3×2×2×4 factorial design was conducted. The design and data features of interest were sample size (50, 300, 1000), number of items per latent variable (3, 5, 7), degree of model misspecification (correctly specified model, misspecified model), nature of the relationships between items and latent variables in the measurement models (reflective,

formative), and the four estimation methods named. Rate of convergence, bias of goodness of fit and estimates of model parameters and standard errors, and accuracy of standard error estimates were evaluated to determine the ability of each estimation method to recover model estimates under each experimental condition. The results indicate that when applied to normally distributed data, Maximum Likelihood generally outperforms the other three estimation methods across experimental conditions. The present study used simulated data to evaluate the performance of four estimation methods when applied to relatively simple structural equation models with small samples and normally distributed data, but future research will need to evaluate the performance of these methods with more complex models and data that is not normally distributed.

Acknowledgements

I am lucky to have had an amazing support system throughout my education. From teachers and faculty advisers, to friends and family, my journey has been filled with caring, creative individuals who inspired and supported me along the way. I would like to thank my graduate adviser, Dr. James Bovaird, for always acting in my best interests, but not hovering too closely. I would also like to thank the other members of my supervisory committee, Drs. Charles Ansorge, Jolene Smyth, and Greg Welch, for their support and flexibility in the time before, leading up to, and during my whirlwind dissertation.

I cannot name everyone who has been there for me along the way, but I am grateful to fellow graduate students who kept me laughing through the final countdown. I would not have made it to the end of this journey without the people who shared my excitement, endured my anxiety, tolerated my neurotic text messages, and validated all the emotions I had along the way whether or not they were justified. For helping me battle those 99 luftballons, I offer my heartfelt appreciation to my sister, Cynthia Estep, and fellow graduate student Natalie Koziol.

This most recent journey has been a long and winding adventure. My children, Amelia and Alex, made the crazy trip nothing short of magical. Thank you both for the endless support, giggles, cuddles, and maniacal laughter. You are amazing, unique people who are never afraid to show your true colors. Finally, I wish to thank my mother, Sandra Chumney, for everything she has done to help me reach this goal. Because of you, Mom, I always knew my children were with someone who loved them, and that made it possible to focus on other responsibilities. I appreciate you for always reminding me to slow down, but never trying to break my stride.

Table of Contents

Table of Tables

Table of Figures

CHAPTER I. INTRODUCTION

In response to increasing expectations from funding agencies, trends in educational research require scientists to investigate increasingly complex phenomena with regard for the contexts in which they occur. These additional layers of exploration and understanding lead to increasingly complex hypotheses and require advanced statistical techniques. Structural equation modeling (SEM) is a common analytic approach for dealing with complex systems of information. Despite their flexibility (Zhu, Walter, Rosenbaum, Russell, & Raina, 2006), traditional SEM methods require large samples in general, and even larger samples for estimating complex models. For applied researchers, large samples are often difficult and sometimes impossible to obtain.

Consider, for example, a recent mail survey of elementary-level teachers which had as its purpose the evaluation of professional development experiences related to four specific areas of academic content and instructional decision-making (i.e., science, reading, math, data-based decision making; Glover, Nugent, Sheridan, Bovaird, & Chumney, 2013). In addition to the typical response rate challenges posed by mail surveys, this particular study was further limited in that fewer than half of all respondents had participated in professional development directly tied to one of the four areas of interest. One goal of the research was to evaluate differences in those professional development experiences between teachers serving at schools located in rural *vs.* non-rural geographic settings. It was necessary for the researchers to break down the sample of participants who had participated in an appropriately-focused professional development experience into smaller subgroups based on the content area focus of their

training and geographic locale. As a result, a typically satisfactory sample size quickly diminished.

A second scenario addresses a context in which large samples are not possible regardless of the resources available to potentially increase sample size or target a specific population precisely. Educational policy makers are often interested in evaluating student academic performance within a single state for the purposes of allocating resources to public schools, comparing the quality of education across school districts/regions, and/or evaluating the performance of teachers and academic administrators. Despite having access to every child in every school district, such research often struggles with the issue of small samples because the population of students within districts – particularly rural districts – is often quite small. In the case of individual teacher evaluation, this sometimes means that data for only a handful of students can be collected. Situations such as these are not uncommon in fields such as education and the social sciences. Unfortunately, traditional SEM techniques are not equipped to handle these types of challenges.

The most common estimation method used with SEM is maximum likelihood (ML; Hoyle, 2000). ML has been studied across myriad contexts and data conditions, and its limitations are well documented. One context in which ML does not perform well is in the presence of small samples (Kline, 2011). Due to this limitation, it is imperative that researchers investigate the utility of alternative approaches to recovering parameter estimates (e.g., partial least squares (PLS), generalized structural components analysis (GSCA), Markov Chain Monte Carlo (MCMC)). If the strengths and weaknesses of each

alternative method in the context of small sample research were more fully understood, researchers would be better equipped to make informed decisions with regard to selecting appropriate estimation methods and interpreting results.

As the field of methodology has advanced, alternative estimation methods have developed and include generalized least squares, weighted least squares, PLS, GSCA, and MCMC approaches. Unfortunately, the performance of these alternatives is not well understood, and their performance with real data is often difficult to predict (Henseler, 2012; Hwang, Ho, & Lee, 2010; Hwang Malhotra, Kim, Tomiuk, & Hong, 2010). Although estimation methods other than those described here have been developed for use with SEMs when the assumptions of ML are violated (e.g., robust ML, weighted least squares), it is not feasible to compare and evaluate the performance of all such alternatives in a single study. Thus, the present study will focus solely on the differential performance of ML, PLS, GSCA, and MCMC methods because they represent diverse and promising approaches for addressing the problem of estimating SEMs with small samples.

Approaches to SEM estimation may be described as covariance-based (e.g., ML) and component-based (e.g., PLS, GSCA), or as frequentist (e.g., ML, PLS, GSCA) and Bayesian (e.g., MCMC). Covariance-based approaches to SEM are designed for model evaluation and validation, while component-based approaches are intended for score computation and prediction (Tenenhaus, 2008). Simply put, the primary distinction between covariance- and component-based estimation is that the former is suited to model testing and the latter is better suited to explaining variance and making predictions

(Hulland, Ryan, & Rayner, 2010; Tenenhaus, 2008). Frequentist approaches identify parameter values represented by observed data (which may or may not consist of true values), while Bayesian approaches describe parameter estimates as abstract representations of relationships based on observed data. In addition to these differences of purpose and perspective, ML, PLS, GSCA, and MCMC also differ in their robustness to varying data conditions, including sample size, number of items, model misspecification, and type of indicator-latent variable relationship (i.e., reflective *vs.* formative measurement models).

Inherent to traditional estimation methods (i.e., ML) is the expectation of large samples. Specifically, the parameter estimates produced by ML are based on asymptotic theory, which implies large samples (Tanaka, 1987). Therefore, as sample size decreases, methods such as ML do not perform as well (e.g., Lee & Song, 2004). Proponents of PLS and GSCA often promote it as performing well in instances of small samples (e.g., Chin & Newsted, 1999; Hulland et al., 2010; Hwang, Ho, et al., 2010; Hwang & Takane, 2004), but both methods have been found to perform inconsistently at times (e.g., Henseler, 2012; Hwang, Ho, et al., 2010; Hwang, Malhotra, et al., 2010), which indicates that more work is needed to understand the interactions between sample size and other design features. Similarly, MCMC implemented as an estimation method within the framework of Bayesian analysis is often viewed as a viable alternative to ML because its sampling procedures make estimation with small samples more feasible, but this approach also does not perform consistently across all combinations of models and sample sizes (e.g., Lee & Song, 2004).

Just as the performance of estimation methods is expected to improve with increased sample size, estimation methods are expected to produce more reliable parameter estimates as the number of items per latent factor increases (e.g., Boomsma, 1982; Velicer & Fava, 1998). As illustrated by Marsh, Hau, Balla, and Grayson (1998), however, increasing the number of items does not necessarily improve the ability of an estimation method to recover parameter estimates. The relationship between quality of parameter estimates and number of items per latent variable has not been studied at length in the context of PLS or GSCA.

In both substantive and methodological research endeavors that utilize SEM, inferences and conclusions are the result of the model used. Although it is difficult to know whether or not theoretical models are specified correctly in applied research, simulation-based research has illustrated the impact of misspecification on parameter recovery across estimation methods (e.g., Asparouhov & Muthén, 2010; Hwang, Malhotra, et al., 2010). The extent to which estimates are impacted by the misspecification of the model depends on design features such as sample size (e.g., Henseler, 2010; Tanaka, 1987) and overall complexity of the model (e.g., Tanaka, 1987).

Whether the relationships between observed variables and latent constructs are formative or reflective in nature is as important to methodological study as it is to theory-driven, applied research. In the context of SEM, latent variables can be modeled as the cause of those observed values (reflective; Bollen & Lennox, 1991), or as a representation of the combined values of those observed values (formative; Curtis & Jackson, 1962). SEMs should be specified to reflect the correct theoretical relationships,

but estimation methods sometimes vary in their performance depending on the type of relationship specified. Until recent years, it was held that SEMs including formative measurement models were inappropriate for traditional ML approaches altogether (Chin, 1998; Ringle, Götz, Wetzels, & Wilson, 2009). More recently, it has been found that ML is likely to overestimate parameters in formative measurement models and underestimate parameters in reflective models (Ringle et al.) when the sample is not large. In contrast to ML, Ringle et al. found that PLS is likely to underestimate parameters in formative models and overestimate parameters in reflective models. The flexibility of GSCA to handle either reflective or formative items has been documented, but the claim is generally based on theoretically-driven expectations of the method without the benefit of empirical evidence (e.g., Hwang & Takane, 2004).

Although some work exists comparing ML to MCMC in a Bayesian framework (e.g., Browne & Draper, 2006) and PLS to GSCA (e.g., Tenenhaus, 2008), the four methods have only been compared once. Chumney (2012) investigated the application of PLS, GSCA, and MCMC to a substantive data set for the purpose of validating the parameter estimates recovered using ML with a small sample and multiple groups. Few consistent patterns of relative bias (i.e., a single estimation method consistently overestimating or underestimating path coefficients, relative to those recovered by ML) of parameter estimates emerged when PLS, GSCA, and MCMC were compared to the ML results. This work identified a gap in the existing literature, as no explanation for the varying performance of the methods was identified. Further, because these data were part of an applied research project, the true population parameters for the specified model

were unknown, and any attempt at explaining the inconsistencies in the performance of the four estimation methods based on those findings would constitute nothing more than conjecture. This is but one example of the extent to which PLS, GSCA, and MCMC approaches are not understood, as researchers are sometimes unable to correctly predict the performance of these methods even in the context of simulation research. For the purpose of contributing to the current understanding of these methods, this study will constitute a systematic evaluation of ML, PLS, GSCA, and MCMC under varying data conditions common to applied research.

Present Study

The present study is a first attempt to compare the relative performance of ML, PLS, GSCA, and MCMC simultaneously under sub-ideal data conditions. Researchers have previously compared different combinations of these approaches under some data conditions, but this is the first known attempt to examine the four methods in a single study. The overarching goal of this study is to understand the effects of sample size, number of items per latent variable, model misspecification, and the nature of the latent variable-indicator relationships on the performance of ML, PLS, GSCA, and MCMC. To guide the process by which this goal will be reached, four specific research questions are posed:

1. To what extent does sample size affect the relative ability of the estimation methods to estimate global model fit and accurately recover model parameters (i.e., item loadings for the measurement model and regression coefficients in the structural model) and their standard errors? It is hypothesized that ML, PLS, and

MCMC will perform better with the larger sample size, regardless of whether the model is correctly specified. It is further hypothesized that ML will produce more biased parameter estimates and less efficient standard error estimates as sample size decreases, compared to PLS, GSCA, and MCMC.

2. To what extent does the number of items per latent variable affect the relative ability of the estimation methods to estimate global model fit and accurately recover model parameters and their standard errors? It is hypothesized that all four estimation methods will perform better with fewer items per latent variable.
3. To what extent does model misspecification (i.e., exclusion of cross-loadings that exist in the population model) affect the relative ability of the estimation methods to estimate global model fit and accurately recover model parameters and their standard errors? It is hypothesized that GSCA will produce more efficient estimates of standard errors than ML or PLS when the model is misspecified. It is also hypothesized that PLS will perform better under conditions of correct specification compared to misspecification.
4. To what extent does the nature of the latent variable-indicator relationship affect the relative ability of the estimation methods to estimate global model fit and accurately recover model parameters and their standard errors?

CHAPTER II. LITERATURE REVIEW

Structural equation modeling is a method for examining a set of relationships and assigning a quantitative value to each based on the covariances among the variables. These quantitative values, referred to as parameter estimates, are numeric approximations of the strength and direction of inter-variable relationships that might be observed in the population (Bollen, 1989; Kline, 2011). A common approach across myriad disciplines (e.g., education, psychology, sociology, economics, marketing research; Monecke & Leisch, 2012), SEM is essentially the concurrent calculation of multiple regression coefficients for a system in which predictor and criterion variables are expected to be interrelated in potentially complex ways (e.g., some variables are both criterion and predictor variables, some criterion variables have multiple predictors, etc.; Bollen, 1989; Haenlein & Kaplan, 2004; Kline, 2011). SEM has the goal of identifying a single set of parameter estimates (i.e., path coefficients, error terms, etc.) that minimizes the total difference between the covariances implied by the model and those observed in the population. SEM is generally comprised of a measurement model(s) and a structural model (Bollen, 1989; Kline, 2011).

The measurement model (sometimes referred to as the outer model; Ringle et al., 2009) connects each latent variable to the observed variables with which it is associated, thereby specifying the synthesis of multiple variables into composite (and sometimes latent) variables. The structural model (also known as the inner model; Ringle et al., 2009) connects the composite (latent) variables within a model to each other. A computational procedure, often referred to as an estimation method, is necessary to

estimate the values of the parameters that describe those relationships. In the SEM context, both the predictor and outcome variables may be latent or observed (Lee & Xia, 2008).

Model Estimation

The process of specifying a model for a given data set and obtaining estimates of the parameter values is called model estimation. Simply put, an estimation method is the method used to reach a set of estimates for a model, an estimator is a particular statistic of interest used to approximate a population parameter (e.g., mean, standard error, path coefficient), and an estimate is the actual value produced for an estimator by the given method of estimation (Kline, 2011).

Several estimation methods and variations of those methods have been developed and applied to SEMs, including maximum likelihood (ML), and ML with robust standard errors (MLR; Muthén & Muthén, 1998-2010), generalized least squares (GLS), and weighted least squares (WLS). However, all of these methods are known to perform poorly under some conditions. Specifically, ML and WLS typically fail to produce accurate parameter estimates when applied to small samples (e.g., ML; Hoogland & Boomsma, 1998; Hu, Bentler, & Kano, 1992; Olsson, Foss, Troye, & Howell, 2000); the more precise estimates produced by MLR are generally restricted to estimates of standard errors instead of path coefficients; GLS is generally insensitive to model misspecification, which leads to overly confident fit statistics (i.e., inflated Type I error; Olsson, Troye, & Howell, 1999). In response to the limitations of these and other similar estimation methods, additional estimation approaches have been applied to the estimation

of SEMs, including partial least squares (PLS; Wold, 1975), generalized structured component analysis (GSCA; Hwang & Takane, 2004; Kline, 2011), and Markov Chain Monte Carlo (MCMC; Hastings, 1970). These three estimation methods and ML are the focus of the present study.

Maximum Likelihood

ML is an estimation method that attempts to minimize the differences between observed data and an imposed model, thereby maximizing the likelihood that the observed data come from a population consistent with the implied model (Kline, 2011). ML is a full-information method that uses an iterative process to obtain the best possible estimates before reaching the convergence criterion. In this context, the "best" possible estimates are those that lead to minimal (or no) differences between estimates produced by subsequent iterations, thereby optimizing the fit function. The fit function of an estimation method is the statistical criterion the method aims to minimize; in ML, the fit function is the difference in covariance structures between the observed data and the population data specified by the model being estimated. The ML fit function is represented as

$$F_{ML}(\hat{\theta}) = \log|\Sigma(\hat{\theta})| + tr\left(S\Sigma^{-1}(\hat{\theta})\right) - \log|S| - (p+q) \quad (1)$$

where $\Sigma(\hat{\theta})$ is the covariance structure, $\hat{\theta}$ are estimated parameters, *tr* is the trace of a matrix, *S* is the covariance matrix observed in the data, Σ^{-1} is the inverse of a matrix, *p* is the number of observed indicators for the endogenous latent factors, and *q* is the number of observed indicators for the exogenous latent factors (Bollen, 1989).

ML is one of the most common and widely used methods for estimating SEMs, is available within SEM software, and yields accurate parameter estimates when used correctly (Kline, 2011). Other advantages of ML are that it is scale free (standardized parameter estimates will not change when a variable is transformed linearly) and scale invariant (the fit function is independent of the scale of response data). Inherent to the use of ML are its assumptions, which include multivariate normality, complete data, and large samples (Bollen, 1989; Kline, 2011). ML is typically the preferred method of estimation within the SEM context because it yields unbiased, consistent, and efficient parameter estimators when its assumptions are satisfied (Bollen, 1989). Despite the availability of literature addressing the importance of meeting these assumptions, the consequences of violating them are not fully understood by all researchers who utilize the method. Thus, ML is often applied in situations where these assumptions are violated, and the result can be biased (i.e., consistently overestimated or underestimated) parameter estimates and standard errors, even when the model is correctly specified (Gerbing & Anderson, 1985; Hwang et al., 2010).

On the one hand, ML is a powerful tool when used correctly, and some research has shown that it is robust to some violations of its assumptions (e.g., Babakus, Ferguson, & Jöreskog, 1987; Maas & Hox, 2004). On the other hand, the fairly stringent assumptions imposed by ML often make it an inappropriate estimation method when used in the context of real-world data characterized by small samples, unknown population models, and other sub-ideal conditions. Specifically, ML relies on asymptotic theory, which implies large samples and assumes correct model specification,

independent observations, independent exogenous variables (i.e., values obtained for exogenous variables are independent), and that the conditional distribution of scores for endogenous variables in the population is multivariate normal (Kline, 2011). Speaking generally, a small sample is problematic in the context of ML because the estimates and fit tests it produces are not asymptotically true (Lee & Song, 2004). This means that without large samples, the validity of statistical inferences may be rightly questioned. ML is known to be robust to minor violations of its assumptions, but the extent of that robustness varies with the data and model.

Partial Least Squares

PLS is a component- (variance-) based approach to modeling developed by Wold (1975) as an alternative to covariance-based estimation methods. Compared to traditional approaches to SEM (i.e., ML), PLS is a more flexible approach that aims to maximize the amount of variance in the dependent variables that is explained by the independent variables (Haenlein & Kaplan, 2004; Wold, 1975). PLS is particularly well suited for small samples (Chin & Newsted, 1999; Haenlein & Kaplan, 2004; Hulland et al., 2010), instances in which large numbers of indicators are used to measure latent constructs (Chin & Newsted, 1999; Haenlein & Kaplan, 2004), cases in which formative indicators serve as the primary source of direct measurement (Fornell & Bookstein, 1982; MacCallum & Browne, 1993), situations in which data are characterized by skewed distributions (Bagozzi & Yi, 1994), and structural model misspecification (Cassell, Hackl, & Westlund, 1999).

Whereas covariance-based approaches to SEM estimate model parameters first, PLS first estimates the latent variable values as the product of linear combinations of indicators (Haenlin & Kaplan, 2004). Another important distinction between ML and PLS in the context of applied research is that ML is likely to produce estimates that are more statistically accurate, but PLS estimates are often more accurate in the prediction of future values (Vinzi, Trinchera, & Amato, 2010). Both listwise deletion and mean imputation are viable options for handling missing data in most PLS software packages (Temme, Kreis, & Hildebrandt, 2006; Tenenhaus, Vinzi, Chatelin, & Lauro, 2005).

PLS estimates are obtained as the result of an iterative five-step process (Henseler, 2010; Tenenhaus, 2008) during which subparts of the overall model are estimated sequentially. It is the simplicity of the approach of sequential regression analyses that allows PLS to be used with small samples; because parameters are estimated individually or in blocks, the complexities of the model are not taken into account simultaneously so larger samples are not necessary (e.g., Reinartz, Haenlein, & Henseler, 2009). The five steps included in the process of PLS during which both the measurement (outer) and structural (inner) model parameter values are estimated are completed as follows:

Step 1: Each latent variable is grouped with its indicators to create blocks of variables and relationships.

Step 2: Outer approximations of the latent variable scores are calculated as linear combinations of the indicators associated with each latent variable,

$$\eta = w_1 x_1 + w_2 x_2 + \cdots + w_p x_p \quad (2)$$

where η is a latent variable, x_1 - x_p are manifest variables associated with that latent variable (regardless of whether the model specifies this portion of measurement to be reflective or formative), and w_1 - w_p are weights assigned to those indicators.

Step 3: Inner weights (w) are calculated to reflect how strongly a latent variable relates to other latent variables in the model; three methods are available for the estimation of inner weights: centroid, factor weighting, and path weighting (Henseler, 2010; Monecke & Leisch, 2012; Tenenhaus, 2008). The centroid method estimates the inner weights based on the signs of the correlations between a latent variable and its adjacent latent variables. The factor weighting method estimates the inner weights based on combinations of correlations between a latent variable and its adjacent latent variables. The path weighting method estimates inner weights based on the directions of the arrows linking latent variables in the model.

Step 4: Inner approximations of latent variable scores are calculated as linear combinations of the outer approximations of the latent variable scores (values obtained in step 2).

Step 5: Estimations of outer weights are calculated based on the relationships between each latent variable and its indicators. In the case of reflective indicators, outer weights are calculated as the covariance between the indicators and the inner approximations of latent variable scores obtained in step 4 (this method is known as Mode A). In the case of formative indicators,

outer weights are calculated as a function of the regression weights obtained from OLS regressions of the inner approximations of latent variable scores (step 4) on the indicators associated with the latent variable (Mode B).

Steps 2-5 are iterative until the change in the outer weight estimates meets a change criterion, at which time step 2 is repeated and latent variable scores for all latent variables are obtained and individual case values are calculated as

$$\eta_1^{"} = w_2\eta_2^{'} + \cdots + w_i\eta_j^{'} + w_x\xi_1^{'} + \cdots + w_i\xi_i^{'} \quad (3)$$

where w_1 - w_p are weights obtained during step 3, η are latent endogenous variable estimates (step 4), and ξ are latent exogenous variables estimates (step 4).

PLS is often viewed as more appropriate for exploratory work than for confirmatory modeling, as its resulting coefficients are generally consistent but biased compared to other estimation methods (Cassell et al., 1999; Lohmöller, 1989). Specifically, in applications of data characterized by both a small sample and a small number of indicators per latent variable, Dijkstra (1983) reported that PLS underestimated the correlations between latent variables (the structural model) and overestimated factor loadings (the measurement model).

The primary advantage of PLS over covariance-based estimation methods such as ML is that it relies on ordinary least squares (OLS) regression to obtain parameter estimates (Jöreskog & Wold, 1982; Wold, 1982) and bootstrap resampling to create standard errors (Monecke & Leisch, 2012), thus relieving the challenge of strong distributional assumptions (Bagozzi & Yi, 1994; Fornell & Bookstein, 1982; Hwang & Takane, 2004; Wold, 1982). PLS is especially flexible, as it can be applied to all data

regardless of measurement scale (Haenlein & Kaplan, 2004). Cassel et al. (1999) demonstrated the robustness of PLS to models that include skewed or multicollinear indicators and some minor structural model misspecification. An additional advantage of PLS is that it is not known to converge to improper solutions (Fornell & Bookstein, 1982; Hanafi, 2007).

The primary disadvantage of PLS is that it does not work toward the minimization of a global optimization criterion (i.e., a fit function; McDonald, 1996), and because of this, there is no meaningful way to define how PLS models are optimized. Thus, an overall goodness of fit statistic is not available for PLS models, which makes it difficult to evaluate the performance of this estimation method (Hwang & Takane, 2004; McDonald, 1996). Tenenhaus et al. (2005) proposed a method for evaluating PLS model fit based on the communality of the measurement model estimates and the redundancy of the estimates of the structural model (discussed later). A modified approach to communality and redundancy has also been developed (presented in Tenenhaus et al., 2005), but is beyond the scope of this paper. An alternative (and much more common) method for evaluating the performance of PLS has been to focus on the recovery of regression coefficients within the structural model (e.g., Vinzi et al., 2010).

Despite its lack of assumptions and being further developed to handle more complex modeling issues in recent years, PLS is not understood well enough for researchers to correctly and consistently predict its performance. For instance, Hwang, Malhotra, et al. (2010) reported that PLS produces more accurate standard error estimates than ML under conditions of model misspecification, but that ML outperforms PLS in

this regard when the model is correctly specified. Hwang et al. reported that PLS performed as well as GSCA, but only when the model was specified incorrectly to exclude cross-loadings; when the model was specified correctly and included cross-loadings, PLS did not perform as well as either ML or GSCA. However, under conditions of correct model specification, PLS produced unbiased estimates of standard errors associated with the parameters of the measurement model, but the standard error estimates for the structural model were found to be biased. These are important findings, as they violated the researchers' expectations and demonstrated the need for additional work using PLS so that the contexts in which it performs reliably might be better understood.

Generalized Structured Component Analysis

GSCA was developed as an alternative to covariance-based methods for SEM and in response to the primary disadvantage of PLS. Specifically, GSCA is a component-based estimation method that was developed in such a way that an overall measure of model fit is available (Hwang & Takane, 2004). The general estimation process for GSCA is the same as PLS, except that GSCA utilizes a fit function which aims to maximize the average amount of explained variance for linear composites of latent variables (Henseler, 2012) and estimates the measurement and structural models simultaneously. Despite its relative newness to the field (introduced in 2004), GSCA has been extended to accommodate higher-order components (Hwang & Takane, 2004), fuzzy clustering (Hwang, DeSarbo, & Takane, 2007) and multicollinearity (regularized model; Hwang, 2009).

The GSCA approach is made up of a method for specifying models, an optimization criterion, and an algorithm used to calculate parameter estimates (Henseler, 2012). GSCA combines the observed variables' values to form linear composites under the assumption that the observed data have been standardized (Hwang & Takane, 2004). Latent variables are calculated as

$$\eta_i = W' z_i \tag{4}$$

where η is a vector of latent variables for respondent i, W is a matrix of component weights associated with the observed variables, and z is a vector of responses for respondent i. These composites are further defined in terms of the relationships between the observed variables and the latent variables. When the model includes formative constructs, GSCA assumes no measurement error in the observed data and the observed values are simply combined in a linear fashion. In the case of reflective constructs, each observed variable is transformed into its own composite, which includes the unit weight. The GSCA measurement model is calculated as

$$z_i = C' \eta_i + \varepsilon_i \tag{5}$$

where C is a loading matrix for the relationships between the latent and observed variables, and ε is a vector of residuals associated with respondent i's observed variable responses. The GSCA structural model is calculated as

$$\eta_i = B' \eta_i + \xi_i \tag{6}$$

where B is a matrix of path coefficients describing the relationships between the latent variables, and ξ is a vector of residuals associated with respondent i's latent variable scores.

The algorithm at work in GSCA is an alternating least squares (ALS; de Leeuw, Young, & Takane, 1976) approach that involves an iterative process by which *A* (a matrix of the relationship between component loadings and their observed variables) is updated for fixed points *V* and *W* (matrices of component weights for the endogenous and exogenous variables, respectively), and then *V* and *W* are updated for fixed point *A*. The optimization criterion of GSCA attempts to minimize the sum of squares of all residuals (Hwang et al., 2010); the fit function can be specified as

$$f_{GSCA} \equiv SS(E) = SS(ZW - ZWA) \quad (7)$$

where *S* is the observed correlation matrix, *Z* is the data matrix composed of the number of observations × number of observed variables, *W* is a matrix of measurement weights, and *A* is a matrix of component loadings and path coefficients (Henseler, 2012).

The advantages of GSCA are similar to those of PLS, in that it is not known to converge to improper solutions, produces unique component score estimates, is not burdened by strict distributional assumptions (Henseler, 2012; Hwang & Takane, 2004), and outperforms ML when models are misspecified (Hwang, Malhotra, et al., 2010). Like PLS, GSCA utilizes bootstrap resampling to estimate standard errors of parameter estimates. An additional advantage of GSCA is that it appears to perform well when applied to both large and small samples (Hwang, Malhotra, et al., 2010; Hwang & Takane, 2004). Compared to PLS, GSCA has the further advantage of being able to estimate multiple group models with equality constraints across groups (Hwang & Takane, 2004).

A noteworthy disadvantage of GSCA is that, despite its positive performance under conditions of model misspecification, GSCA sometimes is outperformed by ML when the model is correctly specified, even when the sample is small (e.g., Henseler, 2010). The primary disadvantage of GSCA is that, as a relatively new estimation method, extensive research on its flexibility has not been conducted. For instance, a method for applying GSCA to models that include interactions among latent variables was only introduced in 2010 (Hwang, Malhotra, et al., 2010), and an application of GSCA to (fuzzy) clustered response sets was introduced in 2007 (Hwang et al., 2007), but neither application has been examined comprehensively.

GSCA is a compromise between principal components analysis and ordinary least squares regression. Like PLS, GSCA utilizes a component-based approach to SEM estimation (Tenenhaus, 2008), but in GSCA, the components used for analysis are linear combinations of the model's observed variables. Compared to both ML and PLS, GSCA has been found to be more robust to model misspecification, and to produce more precise estimates of standard errors regardless of whether the model is correctly specified (Hwang, Malhotra, et al., 2010). Because it does not impose distributional assumptions, GSCA is often touted as a viable alternative to ML estimation with small samples. This is supported by Hwang, Malhotra, et al., who reported that GSCA provided more accurate standard error estimates than either ML or PLS regardless of sample size. However, there is still a lot unknown about GSCA and its performance under varying data conditions, including small samples (e.g., Henseler, 2012; Hwang, Malhotra, et al., 2010).

Markov Chain Monte Carlo

A Markov chain is a series (or chain) of samplings from a distribution for which the probability of each successive sample is dependent on previously sampled values, given the most recent value (Carlin & Chib, 1995; Geyer, 1992; Lynch, 2007). This iterative process can be expressed

$$P(x_{N+1}|x_0, \cdots, x_N) = P(x_{N+1}|x_N)$$

where x_i (i = 0,...,N) represents the number of iterations. MCMC is the use of Markov chain sampling as the method for estimating parameter values (Geyer, 1992; Lynch, 2007). In the context of SEM, the MCMC algorithm can be applied to either frequentist or Bayesian approaches (Cowles & Carlin, 1996). For the purposes of the present study, MCMC is discussed here only as it functions within the context of Bayesian model estimation.

Bayesian estimation differs from frequentist approaches (e.g., ML, PLS, GSCA) with regard to what it is that is being estimated. Whereas frequentist estimation methods view parameters as constants and work to identify the estimates for those parameters that produce the best model-data fit, Bayesian methods view parameters as random variables and work to combine the likelihood of the data with prior distributions to form posterior distributions from which to draw plausible values for the parameter estimates (Muthén, 2010; Muthén & Asparouhov, 2011). In other words, the frequentist perspective holds that true population parameters exist but can only be determined through data, and the Bayesian perspective posits that population parameters are abstract explanations of the relationships between data.

The basic process of the Bayesian approach to model estimation is to create a prior distribution of possible values from which to sample a value, combine that sample with the likelihood of the value given the data to create a posterior distribution, and then use the posterior distribution to update the prior distribution (Lynch, 2007; Muthén, 2010). This iterative process is completed for each parameter being estimated. The Bayesian approach is based on Bayes' Theorem, which can be represented as

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \tag{8}$$

where *A* and *B* are events with joint probability expressed as a function of the conditional and marginal probabilities

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A) \tag{9}$$

The initial prior distribution can be created using either informative or noninformative values. In the instance of informative priors, the researchers' expected values for the parameter estimates (based on theory or past research) are used as the basis for the prior distributions (Lynch, 2007). The posterior distribution, then, is dependent on these starting values. Using informative priors can be advantageous, as they can reduce the amount of time required for the model to converge and result in more accurate estimates (Lee & Song, 2004), as such estimates are expected to be closer to the final answer than a random start value. In the instance of noninformative priors, the researcher may have little or no basis for determining expected values for the parameter estimates. In such cases, random values in the prior distribution may be left equal to zero or chosen such that the prior distribution reflects a uniform distribution. In this case, the prior distribution has little impact on the posterior distribution (Lynch, 2007). A prior

distribution with non-informative priors is sometimes referred to as a beta distribution, and has the probability density function of

$$f(K|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} K^{\alpha-1}(1-K)^{\beta-1} \quad (10)$$

where *K* is the proportion of events which occur to maximize the probability of attaining a given outcome, α and β represents prior values, and *K*, α, and β are random variables (Lynch, 2007). Regardless of the amount of information used to create a prior distribution, the posterior distribution takes the form

$$p(\theta|x) = \frac{p(x|\theta)\times p(\theta)}{p(x)} \quad (11)$$

where $p(\theta|x)$ is the posterior distribution, $p(x|\theta)$ is the likelihood of the data (or, the data given the parameters), $p(\theta)$ is a prior distribution, and $p(x)$ is the observed data. Each parameter estimate obtained via this approach is then a summary of the posterior distribution, typically in the form of its mean, median, or mode (Muthén & Asparouhov, 2011). In the case of MCMC, the Bayesian estimate represents the mean of the posterior distribution (Lee & Song, 2004). Regardless of whether informative or non-informative priors are specified by the researchers, MCMC attempts to work from starting values more appropriate to the data than random values. To do this, a portion of the draws in each MCMC chain are discarded and the values at the end of that portion of the chain are used as starting values for obtaining estimates. This process is known as the burn-in phase, and can be lengthened to improve starting values (e.g., Meyn & Tweedie, 1993).

An advantage of MCMC estimation over ML is that the Markov chain sampling approach does not rely on the assumptions of asymptotic theory, which means that a large sample size is not necessary for drawing valid statistical inferences (Lee & Song, 2004;

Song & Lee, 2006). The Bayesian estimates derived through the MCMC process are not affected by the size of the sample, as they are sampled from the posterior distribution which includes sufficient observations (Song & Lee, 2006). However, although Browne and Draper (2006) demonstrated that Bayesian estimation yields similar results to ML when applied to large samples, they also reported that the method is not robust to small samples under all conditions. Similarly, Lee and Song (2004) reported poorer performance of the MCMC approach with samples fewer than four times the number of parameters in the model, but concluded that MCMC estimation is preferable over ML when the sample size is roughly two or three times the number of model parameters. Specifically, Lee and Song reported that estimates of standard errors were overestimated using this approach. In the context of latent variables, Bayesian estimation is further limited by its relative lack of robustness to model misspecification in the presence of more than a few indicators (Asparouhov & Muthén, 2010). Despite these limitations, MCMC estimation within the Bayesian framework continues to serve as a common alternative to ML.

Each of these estimation methods has distinct advantages and disadvantages. Despite their disadvantages, ML, PLS, GSCA, and MCMC are not uncommon in applied research. Therefore, it is important to investigate the performance of each method under varying data conditions to better understand the extent of their limitations. Through examination of the situations in which these methods perform poorly, it may be possible to discover their relative strengths and the data conditions to which they are robust. ML is included in the current study because it represents the most common estimation approach used in SEM. PLS and GSCA are included in the present study because they represent a

different theoretical approach to the estimation of SEM parameters. Despite their documented strengths, neither method has been studied thoroughly to the extent that researchers fully understand the conditions under which they each perform well or fail to perform to acceptable standards. MCMC is included in the present study because its freedom from distributional assumptions gives it the potential to perform well when applied to small samples. Given that researchers are currently utilizing these estimation methods despite being unable to accurately predict their performance under various data conditions, it is appropriate to conduct research of an empirical nature (i.e., through simulation) to better understand them.

Simulation Research

In the research context, simulation is the practice of generating data to have specific characteristics for the purpose of evaluating those data or the performance of analytic techniques. The advantage of using simulated data over real-world data is that, because the researcher creates the data, he has complete control over the characteristics of the data and the relationships between variables. Knowing the true values of the model used to simulate the data (i.e., the population model) allows the researcher to conduct an empirical evaluation of various analytic methods by comparing the results of various analytic techniques to the certain truths that are known about the data (Paxton, Curran, Bollen, Kirby, & Chen, 2001). Simulation is a common research method used in the study of SEM (e.g., Anderson & Gerbing, 1984; Curran, West, & Finch, 1996; Gerbing & Anderson, 1993; Hu & Bentler, 1999; Hwang et al., 2010), and has been used to study the performance of estimation methods (e.g., Henseler & Chin, 2010; Hwang et al., 2010)

and test statistics and fit indices (e.g., Anderson & Gerbing, 1984; Curran, et al., 1996; Hu & Bentler, 1999), as well as the effects of model and data characteristics such as sample size (e.g., Fan, Thompson, & Wang, 1999; Hox & Maas, 2001) and misspecification (e.g., Hwang et al., 2010).

Present Study

ML, PLS, GSCA, and MCMC estimation methods differ, but each is characterized by its own set of strengths and weaknesses. Generally speaking, the strengths of covariance-based methods (i.e., ML) are the weaknesses of component-based methods (i.e., PLS, GSCA), and the weaknesses of component-based methods are the strengths of covariance-based methods (Jöreskog & Wold, 1982). In essence, these estimation methods should be thought of as complementary techniques, each suited to different purposes and data characteristics (e.g., Hair, Sarstedt, & Mena, 2012). It follows, then, that the choice between estimation methods should be dependent upon the specific goals of the researcher; the selection of either a covariance- or component-based method, or the adoption of either a frequentist or Bayesian perspective should be made with consideration of the model of interest and the data at hand. In practice, this approach is complicated by the fact that very little is known about the relative performance of such estimation methods under varying data conditions.

Purpose Statement

The purpose of this project was to use simulated data to shed some light on this issue by evaluating the performance of ML, PLS, GSCA, and MCMC under conditions not uncommon to applied researchers. Specifically, this research investigated the impact of sample size, measurement model complexity, model misspecification, and the nature of the latent variable-indicator relationships on the relative performance of ML, PLS, GSCA, and MCMC; several hypotheses were generated to guide exploration of the results.

Hypothesis 1: *Compared to PLS, GSCA, and MCMC, ML will result in lower convergence rates across all experimental conditions.* As discussed elsewhere in this paper, ML estimation assumes large samples. Because of this, ML is more likely than other estimation methods to fail to converge when applied to small samples. As alternative methods to ML developed in part to overcome the small sample limitations of ML, PLS, GSCA, and MCMC algorithms are more likely to converge to acceptable solutions.

Hypothesis 2: *ML, PLS, and MCMC perform better as sample size increases, regardless of whether the model is correctly specified.* Despite the ability of ML, PLS, and MCMC to produce accurate parameter estimates under some conditions when samples are small, all three methods have been documented as producing more accurate estimates of parameters and standard errors as sample size increases.

Hypothesis 3: *Compared to PLS, GSCA, and MCMC, ML produces more biased parameter estimates and less biased standard error estimates when the sample size is its smallest (i.e., n = 50).* A basic underlying assumption of ML is that it relies on asymptotic theory which implies large samples (Kline, 2011). As sample size decreased, then, ML was expected to recover less favorable estimates. Because PLS, GSCA, and MCMC do not make the same assumptions, it was expected that they would outperform ML when applied to smaller samples.

Hypothesis 4: *Under conditions of model misspecification, GSCA produces less biased estimates of standard errors than ML, or PLS.* As described by Hwang, Malhotra, et al. (2010) and others, GSCA often produces less biased standard error estimates when

compared to other frequentist approaches under conditions of model misspecification. The conditions of the present study are similar to those utilized in past research; thus, it was expected that this finding would be replicated.

Hypothesis 5: *PLS recovers less biased parameter and standard error estimates for the measurement model when the model is specified correctly compared to when the model is misspecified.* PLS takes a components-based approach to SEM estimation, which means that the measurement and structural models are estimated separately. Because the final estimates for the measurement model are obtained first, those estimates are not additionally influenced by the quality of the final structural model estimates (Haenlin & Kaplan, 2004). In addition, several studies have reported more favorable performance of the PLS approach under conditions of correct model specification compared to model misspecification (e.g., Hwang, Malhotra, et al., 2010).

As the conditions under which each type of estimation method perform best are better understood, applied researchers will become more informed and better equipped to make sound, intentional choices with regard to estimation methods. As practices improve in this way, the inferences that can be drawn will become more meaningful in informing policy development and future research.

CHAPTER III. METHODS AND PROCEDURES

SEM is a common tool in both methodological and applied research endeavors. Traditional approaches to SEM are dependent on covariance-based estimation methods such as ML. More recently, alternative approaches to these methods have been developed, including PLS, GSCA, and MCMC. The primary advantage of these estimation methods is that they are theoretically robust performers in instances when the ideal data conditions are not available, but these methods do not perform to optimum levels under all data conditions. There exists a gap in the literature where an understanding of the factors that impact the relative performance of ML, PLS, GSCA, and MCMC is non-existent. This study will provide a foundational piece for bridging this gap.

Simulation Conditions

The primary goal of this study is to investigate the accuracy with which ML, PLS, and GSCA, and MCMC estimation methods recover the parameters for SEMs under conditions that frequently must be handled in the context of applied research. Specifically, this study examines the impacts of sample size, complexity of the measurement model (i.e., number of items per latent variable), model misspecification, and latent variable-indicator relationships in the context of a relatively simple SEM.

Sample Size

The sample size necessary to yield stable model results is an empirical question that depends on the complexity of the model as well as other contextual factors (e.g., Jackson, 2003). Due to sample size limitations, researchers often apply complex analytic

models to data containing too few cases. The extent to which a diminished sample size impacts research findings under different conditions is not fully understood, with less information available for estimation methods other than ML. Three conditions of sample size were implemented in this study, n = 50, 300, and 1,000. These values were selected to reflect one common rule for sample size in SEM (minimum of 200 cases; Kline, 2011), sample sizes common to research of this type (e.g., Anderson & Gerbing, 1984; Ding, Velicer, & Harlow, 1995; Henseler, 2012; Hwang, Malhotra, et al., 2010; Olsson, Foss, & Breivik, 2004; Paxton et al., 2001), and a large sample intended to demonstrate performance of the estimation methods under a more ideal sample size condition.

Number of Items

The optimal number of items that should be associated with latent variables has been an issue of much study and debate in the SEM literature (e.g., Ding et al., 1995; Guadagnoli & Velicer, 1988; Tomás, Hontangas, & Oliver, 2000; Velicer & Fava, 1998). Based on statistical theory, applied research, and simulation studies, a common rule of thumb is that fewer than three items per latent variable is inadequate (Ding et al., 1995; Tomás et al., 2000). Further, it has been found that power, accuracy, and precision of estimates increases as the number of items per latent variable also increases (e.g., Boomsma, 1982; Guadagnoli & Velicer, 1988; MacCallum, Browne, & Sugawara, 1996; Marsh et al., 1998; Nunnally, 1967; Velicer & Fava, 1998). Despite the number of research studies which have included the number of items per latent variable as a primary variable of interest, the matter is not yet settled due in part to the number of other design characteristics that must be considered, including method of estimation and sample size.

As the number of indicators increases, so does the complexity of the measurement model and the size of the sample necessary to accurately recover parameter estimates.

To date, most empirical investigations into the performance of SEM with varying numbers of items per latent variable have been set within the context of covariance-based modeling methods. Thus, the ability of component-based methods to handle different numbers of items, and the relative performance of covariance- and component- based methods is not yet understood. Three levels of number of items per latent variable were implemented in this study. Specifically, this research investigates the performance of covariance- and component- based estimation methods in the presence of 3, 5, and 7 items. These levels are consistent with both previous simulation research (e.g., Anderson & Gerbing, 1984; Marsh et al., 1998; Velicer & Fava, 1998) and applied analyses (Ding et al., 1995).

Misspecification

Misspecification is a concern for researchers anytime the true model is not known. In instances where the true model is unknown (nearly all applied research endeavors), proper specification depends on a perfect match between the model being evaluated and the theoretical model (Hoogland & Boomsma, 1998). A model can be misspecified in several ways, including the omission of important variables or the inclusion of additional, unnecessary relationships. Model misspecification is problematic for researchers because it does not typically prevent a model from converging and producing estimates, and does not always lead to poor model fit. Thus, the challenge exists because a researcher does not know that parameter estimates may be incorrect if

they are not aware of the misspecification within their model. In cases where the structural portion of a model is misspecified, path coefficients are expected to be biased (Hoogland & Boomsma, 1998). This is not to say, however, that item loadings are also biased.

The extent to which model misspecification impacts parameter estimates depends on the degree of misspecification, contextual effects specific to the constructs and data included in the analysis, and the estimation method. For example, a model which includes misspecification(s) in the structural model that is estimated using a full information method such as ML could result in biased parameters in the measurement portion of the model as well as the structural portion of the model (Hoogland & Boomsma, 1998) due to the simultaneous estimation of the two parts of the model. Limited information estimation methods such as PLS and GSCA may also produce biased estimates under conditions of model misspecification, but because the measurement and structural parts of the model are estimated separately, misspecification in the structural model is not as likely to impact parameter estimates recovered for the measurement model.

Two conditions of model misspecification were implemented in this study: the model was specified correctly or misspecified by excluding cross-loadings that exist in the corresponding population model. Model misspecification is an important variable to consider given its potential to occur in applied research endeavors when the true model is hypothesized or theoretically-based and not known by the researcher (e.g., Hu & Bentler, 1999; Jackson, 2007). The exclusion of cross-loadings is a relatively simple means of

introducing model misspecification, and a common approach in simulation research (e.g., Hu & Bentler, 1999; Hwang, Malhotra, et al., 2010; Paxton et al., 2001).

Latent Variable-Indicator Relationships

Latent variables represent unobservable constructs measured through observable (manifest) variables believed to be related to the latent variable. A reflective latent variable-indicator relationship implies that the latent variable is independent of its indicators and would exist even if its indicators did not (Coltman, Devinney, Midgley, & Venaik, 2008). Examples of reflective latent constructs are personality characteristics such as extraversion and neuroticism – both are characteristics of a person that exist regardless of what personality measure is used to collect data. Because reflective indicators are merely manifestations of a construct that exists without them, it is assumed that a set of reflective indicators are related to each other (Vinzi et al., 2010). This implies that the values of those indicators vary together. For example, as a person becomes more neurotic, it is expected that their values on the reflective indicators associated with that construct will increase. Reflective indicators are related to their latent variable using simple regression (Tenenhaus, 2008). Reflective relationships can be successfully modeled by either traditional covariance-based approaches to SEM or component-based approaches.

A formative latent variable-indicator relationship implies that the latent variable is formed from some combination of its indicators and would not exist without those indicators (Coltman et al., 2008); thus, formative latent constructs are dependent on the indicators used to create them. An example of a formative latent construct is

socioeconomic status (SES), where SES is dependent on the specific items used to measure it (e.g., household income, education, number of parents living in the home, primary language, etc.). In contrast to reflective indicators, a respondent can change levels on one formative indicator without necessarily affecting the other indicators (e.g., household income might increase while education remains constant; Wilcox, Howell, & Breivik, 2008). Formative indicators are related in groups to their latent variables using multiple regression (Tenenhaus, 2008). Formative relationships can be successfully modeled by component-based approaches such as PLS, but are not handled as well by covariance-based approaches (Diamantopolous, 2011).

Correct modeling of latent variable-indicator relationships is important, as the estimation of meaningful relationships in the structural model relies on proper specification of the measurement model(s) (Anderson & Gerbing, 1988; Coltman et al., 2008; Diamantopoulos, Riefler, & Roth, 2008; Roy, Tarafdar, Ragu-Nathan, & Marsillac, 2012). The extent to which proper specification of formative *vs.* reflective relationships impacts parameter estimates is not fully understood (Petter, Straub, & Rai, 2007). Two conditions of latent variable-indicator relationships will be implemented in this study: all relationships will be formative or reflective.

Summary of Experimental Design

The five factors included in this simulation study result in a 3×3×2×2×4 design. The 144 cells of the design represent three sample sizes (50, 300, 1000), three levels of indicators per latent variable (each set containing 3, 5, or 7 items), two degrees of model specification correctness (specified correctly, misspecified), two types of latent variable-

indicator relationships (all relationships formative or reflective), and four estimation methods (ML, PLS, GSCA, MCMC).

Population Models

This study relied on data simulated to reflect SEMs common to applied research as well as other, simulation-based research. Specifically, the present study utilized population models comprised of three latent variables, an equal (but varying) number of items per latent variable, and cross-loadings. The population models and their parameters are similar to those used in previous studies (e.g., Henseler, 2012; Hwang, Malhotra, et al., 2010; Paxton et al., 2001; Tomás et al., 2000).

As noted in previous work by researchers such as Vinzi et al. (2010), the relationship between reflective and formative measurement models is essentially the same relationship that exists between factor models with high reliability among the indicators (reflective models with measurement model error) and factor models with low reliability among the indicators (formative models with essentially no measurement model error; e.g., Diamantopoulos et al., 2008). Thus, for experimental conditions for which formative indicator-latent variable relationships were of interest in the present study, reflective models with low reliability among indicators were used for both data generation and estimating the analytic models. This approach was chosen to allow more consistency across the experimental conditions, as it made it possible for all data sets to be generated in the same manner. Furthermore, conceptualizing the formative conditions as error-free reflective models made it possible to apply each estimation method in the same manner across all data sets (e.g., Mode A estimation was implemented for the PLS

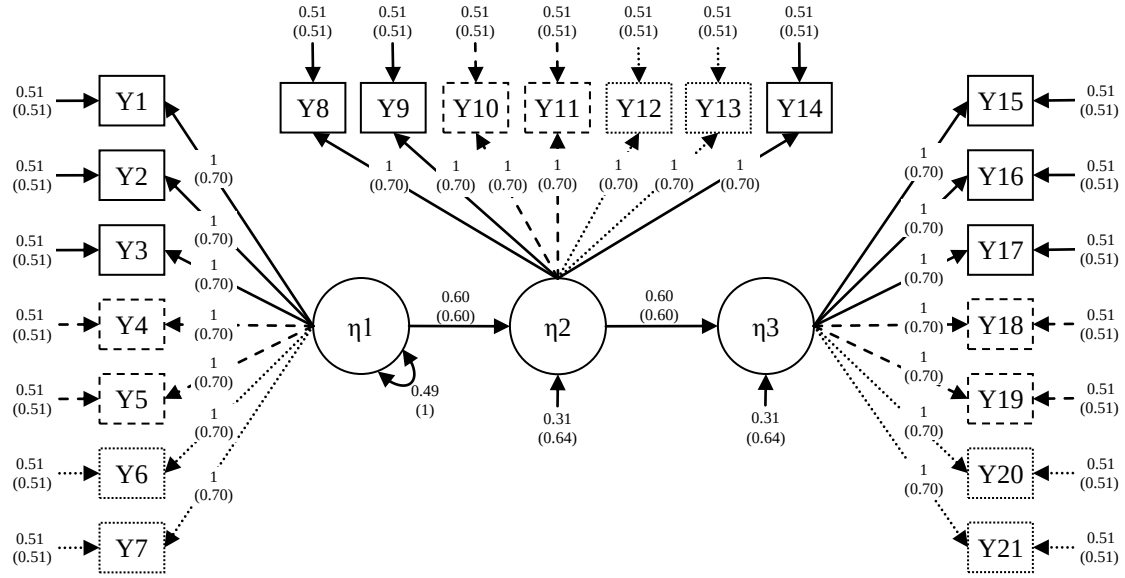
approach across all conditions). Ultimately, this approach allowed differences in the performance of the estimation methods between reflective and formative models in the present study to be attributed to the reflective/formative nature of the measurement model relationships and not to subtle inconsistencies resulting from the estimation methods themselves.

Correct Specification, Reflective Indicators

For the conditions of correct model specification with reflective indicators, the population models specify that all indicator-latent variable relationships have standardized values of 0.700, the error terms associated with all indicators have standardized values of 0.510, the error terms associated with the latent variables have standardized values of 1.000 (η_1) and 0.640 (η_2 and η_3), the path coefficients linking the latent variables each have standardized values of 0.600, and the models do not include cross-loading items (see Figure 1).

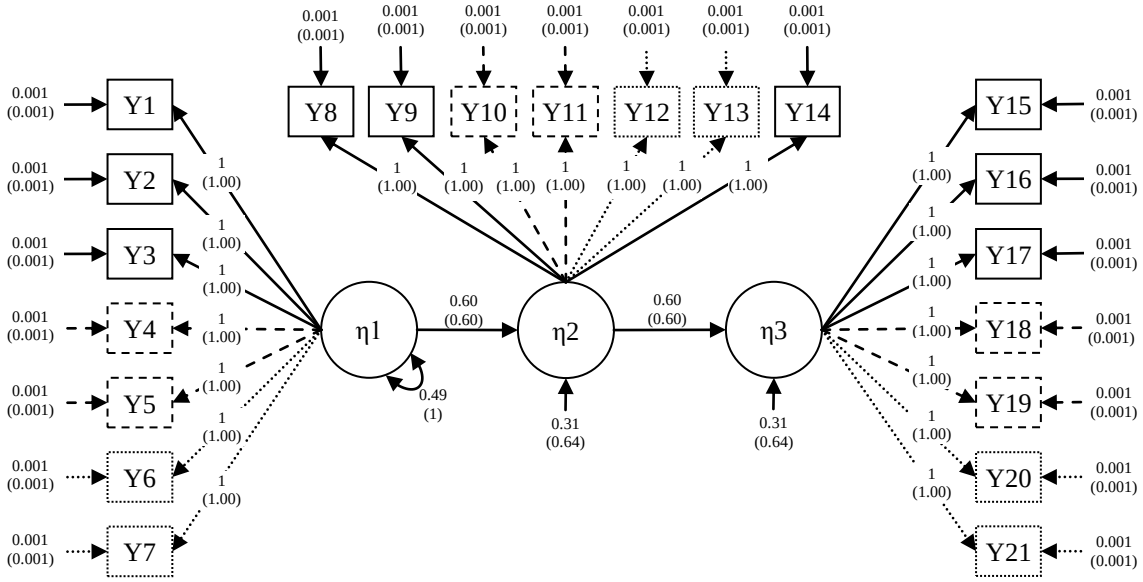
Correct Specification, Formative Indicators

For the conditions of correct model specification with formative indicators, the population models specify that all indicator-latent variable relationships are reflective and have standardized values of 1.000, the error terms associated with all indicators have standardized values of 0.001, the error terms associated with the latent variables have standardized values of 1.000 (η_1) and 0.640 (η_2 and η_3), the path coefficients linking the latent variables each have a standardized value of 0.600, and the models do not include cross-loading items (see Figure 2).



Note: Manifest variables and relationships not indicated by solid lines are included only in 5 (dashed lines) and 7 (dashed and dotted lines) items per latent variable conditions.

Figure 1. Population model for reflective indicators and correct model specification



Note: Manifest variables and relationships not indicated by solid lines are included only in 5 (dashed lines) and 7 (dashed and dotted lines) items per latent variable conditions.

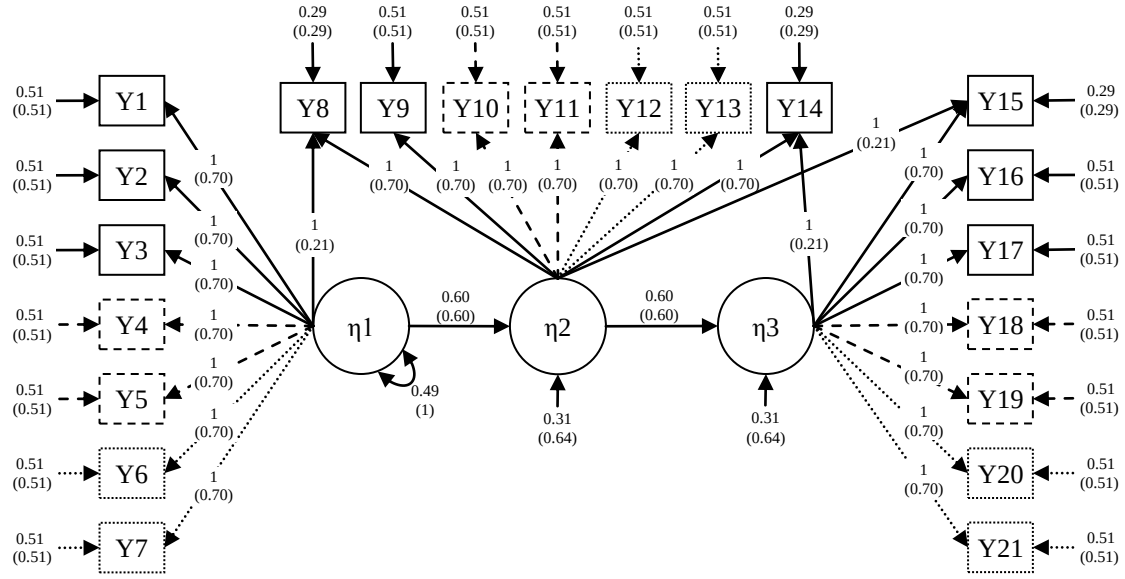
Figure 2. Population mode for formative indicators (reflective relationships with low reliability) and correct model specification.

Misspecification, Reflective Indicators

For the conditions of model misspecification with reflective indicators, the population models specify that all indicator-latent variable relationships have standardized values of 0.700 and 0.210 for items that load on only one latent variable and items that load on more than one latent variable, respectively, the error terms associated with all indicators have standardized values of 0.510 and 0.290 for indicators that load on only one latent variable and indicators that load on more than one latent variable, respectively, the error terms associated with the latent variables have standardized values of 1.000 (η_1) and 0.640 (η_2 and η_3), the path coefficients linking the latent variables each have a standardized value of 0.600, and the models include three items which relate to more than one latent variable (see Figure 3).

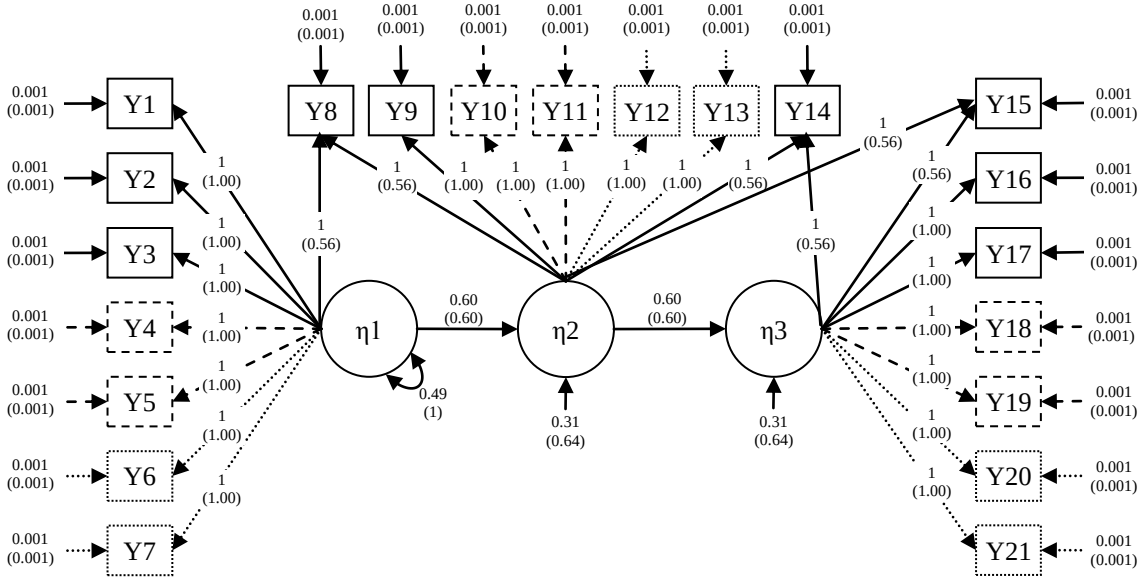
Misspecification, Formative Indicators

For the conditions of model misspecification with formative indicators, the population models specify that all indicator-latent variable relationships are reflective, the indicator-latent variable relationships for items that load on only one latent variable have standardized values of 1.000, all indicator-latent variable relationships for items that load on more than one latent variable have standardized values of 0.560, the error terms associated with all indicators have standardized values of 0.001, the error terms associated with the latent variables have standardized values of 1.000 (η_1) and 0.640 (η_2 and η_3), the path coefficients linking the latent variables each have a standardized value of 0.600, and the models include three items which relate to more than one latent variable (Figure 4).



Note: Manifest variables and relationships not indicated by solid lines are included only in 5 (dashed lines) and 7 (dashed and dotted lines) items per latent variable conditions.

Figure 3. Population model for reflective indicators and model misspecification.



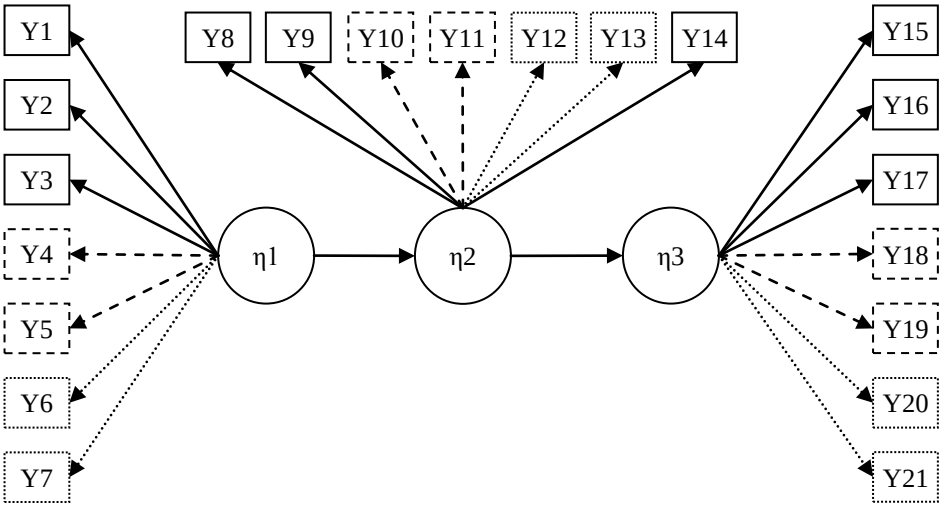
Note: Manifest variables and relationships not indicated by solid lines are included only in 5 (dashed lines) and 7 (dashed and dotted lines) items per latent variable conditions.

Figure 4. Population model for formative indicators (reflective relationships with low reliability) and model misspecification.

Procedures

Using M*plus* (version 6; Muthén & Muthén, 1998-2010), 150 replications of each unique condition were simulated. Recommendations for a sufficient number of replications typically suggest a minimum of 1,000 replications be used (e.g., Hwang , Malhotra, et al., 2010; Muthén & Muthén, 2002). However, due to a software limitation specific to the present study (i.e., the software available for GSCA estimation of models cannot be programmed to complete estimation of multiple data sets in an automatized fashion), estimation of 1,000 replications for the models of interest was not feasible. Although a minimum of 1,000 replications is often recommended, several studies have been completed using these estimation methods which rely on fewer replications and an ANOVA approach to analysis (e.g., 100 replications as reported in Ding et al., 1995; Kankaraš, Vermunt, & Moors, 2011; Lee, Song, & Lee, 2003; Lee & Tang, 2006; Lee & Xia, 2008; Lee & Zhu, 2002; Olsson et al., 2000; Song & Lee, 2002; Song, Lee, & Hser, 2008; and Tomás, Hontangas, & Oliver, 2013, and 200 replications as reported in Fan et al., 1999; Hu & Bentler, 1999; Jackson, 2003, 2007; Lee & Song, 2004; and Song & Lee, 2005, 2006).

The analytic models (as depicted in Figure 5) were fit to each data set using ML, PLS, GSCA, and MCMC. For this step, ML and MCMC estimation were conducted in M*plus* (version 6; Muthén & Muthén, 1998-2010), the plsSEM package (Monecke & Leisch, 2012; refer to Monecke & Leisch for a comparison of plsSEM and SmartPLS parameter recovery performance) developed for R (R Development Core Team, 2012) was used for PLS estimation, and GeSCA was used for GSCA estimation



Note: Manifest variables and relationships not indicated by solid lines are included only in 5 (dashed lines) and 7 (dashed and dotted lines) items per latent variable conditions.

Figure 5. Analytic model for all conditions.

(http://www.sem-gesca.org). For each replication, the maximum number of iterations allowed was set to 1,000, and the number of bootstrap samples used to recover standard error estimates was set to 500 for those methods which relied on bootstrapping to obtain standard error estimates (i.e., PLS, GSCA, MCMC). Finally, convergence rates, global model fit, and the quality of the recovered parameter and standard error estimates were evaluated.

Outcomes of Interest

Evaluating the fit of a model to a particular data set generally consists of some combination of evaluating the fit of the model to the data using available fit indices (test statistics), investigating local model strain by examining the different parts of the model for unnecessary parameters that hurt fit or missing parameters that might improve local fit, and examining model parameter estimates, standard errors, effect sizes, and

significance levels (Brown, 2006; Kline, 2011). In the context of a study such as that presented here, which uses more than one method of estimation, evaluation of the resulting models via fit indices is somewhat complicated by the fact that no one test statistic is generally calculated for the four estimation methods used here. Therefore, evaluation of the results of the present study relied on convergence rates, the Goodness of Fit Index (GOF; Tenenhaus et al., 2005), and evaluation of the recovered parameter and standard error estimates. Together, these methods constitute an appropriate method of model evaluation, given the lack of a comparable test statistic across the four estimation methods and the importance of parameter estimates and their standard errors to the general utility of a model (e.g., Nevitt & Hancock, 2004).

Convergence Rate

Convergence is the point at which an estimation method recovers parameter estimates with a level of precision that meets a predetermined criterion (i.e., the convergence criterion; Fan et al., 1999). In practice, the criteria used to determine convergence is not the same across all estimation methods (Hwang, Malhotra, et al., 2010). Specifically, estimation methods with a fit function have a specific level of increase or decrease in that fit as their convergence criterion; estimation methods which to not have a fit function (e.g., PLS) converge when the change in estimates from one iteration to the next is smaller than some predetermined value and continued iterations are not expected to improve upon the recovered estimates. In cases where a model fails to converge on a solution, parameter and standard error estimates are not produced. Convergence rate was calculated as the proportion of data sets in each condition for

which 1) the estimation method converged, and 2) the recovered estimates consisted only of statistically plausible values (i.e., no negative residual variance estimates; Fan et al., 1999). As suggested by Paxton et al. (2001), only estimates from replications which resulted in converged solutions with plausible values were deemed appropriate for analysis, as the present study was not intended to study the results of improper or non-converged solutions.

Overall Model Fit

The Goodness of Fit Index (GOF; Tenenhaus et al., 2005) was developed as a means of assessing the quality of estimates obtained using PLS estimation. This fit value is calculated from the R^2 values obtained for the structural model and the measurement model by first calculating a communality index (Tenenhaus et al., 2005). For GOF, the communality index for each block (each latent variable and the observed variables to which it relates) is calculated

$$\mathrm{C}_j = \frac{1}{p_j}\sum_{h=1}^{p_j} cor^2\left(x_{jh}, \eta_j\right) \tag{12}$$

where j is a block, p is the number of manifest variables, x is a manifest variable response, and η_j is a component score. The communality index is calculated for each block, and the average communality for the measurement model is calculated

$$\bar{\mathrm{C}} = \frac{1}{p}\sum_{j=1}^{J} p_j \mathrm{C}_j \tag{13}$$

Finally, the global goodness of fit value is calculated as the square root of the mean communality multiplied by the mean of the R^2 values, as

$$GOF = \sqrt{\bar{\mathrm{C}} \times \overline{R^2}} \tag{14}$$

As is obvious from the above formulas, this goodness of fit index is relatively easy to calculate from the computed latent variable scores and R^2 values produced by an estimation model. For example, consider the condition of five reflective items per latent variable. With reference to equation 12, p_j = 5 items per latent variable, x_{jh} = the response of case *h* on item *x*, and η_j = the latent construct score for block *j*. Equation 12 will be calculated for each block of variables in the model, where a block is defined as one latent variable and the items to which it relates (i.e., block 1 consists of η_1, Y1- Y5; block 2 consists of η_2, Y6-Y10; block 3 consists of η_3, Y11-Y15). Thus, for Model 4, equation 12 will be calculated for three blocks (η_1, η_2, η_3). Equation 13, then, is calculated as the average of the values of equation 12 calculated for the 3 blocks. Finally, equation 14 is calculated by obtaining the square root of the product of equation 13 and the mean R^2 value, where the mean R^2 value is the average of the R^2 values obtained for η_1, η_2, η_3. For these calculations, the values of *j*, and p_j are determined by the model; the values of x_{jh}, η_j, and R^2 were provided by the software used for each of the four estimation methods.

Parameter Estimates and Standard Errors

For the purposes of evaluating the ability of the four estimation methods to recover model parameters and their standard errors, the present study analyzed the standardized estimates for all outcomes. Although unstandardized estimates are more commonly used in the evaluation of ML in the context of simulation research, the current available software for both PLS and GSCA provides only standardized estimates.

Parameter Estimates. The quality of the recovered parameter estimates for both the measurement and structural models was evaluated in terms of bias (e.g.,

Hutchinson & Bandalos, 1997). In this context, bias is defined as the proportion of the difference between the sample and population values, relative to the population values (Enders & Bandalos, 2001), and is calculated

$$\%BIAS = \left[\frac{|\theta_i - \theta_B|}{\theta_B}\right] \times 100 \quad (15)$$

where θ_i is the recovered parameter estimate and θ_B is the known population parameter. Average bias was calculated separately for the measurement and structural models in each replication data set.

Standard Errors. The precision of the recovered standard errors associated with the parameter estimates for the measurement and structural models will be evaluated in terms of the mean absolute difference between the standard error estimates and the empirical standard errors (MAD; Hwang, Malhotra, et al., 2010), calculated

$$MAD = \frac{\sum_{j=1}^{P} |\mathrm{SE}(\hat{\theta}_j) - \mathrm{SE}(\theta_j)|}{P} \quad (16)$$

where $\mathrm{SE}(\hat{\theta}_j)$ is the recovered standard error estimate, $\mathrm{SE}(\theta_j)$ is the true value for that standard error, and *P* is the number of parameters. The true values for $\mathrm{SE}(\theta_j)$ were obtained empirically via a Monte Carlo simulation (conducted in M*plus*, version 6, Muthén & Muthén. 1998-2010) which included 500 replications and 2,000 bootstrap resamples per replication for each of 3 (sample size) ×3 (number of items) ×2 (degree of specification) ×2 (high/low reliability) experimental conditions. True (empirical) standard errors were calculated as

$$\mathrm{SE}(\theta_j) = \sqrt{\frac{\sum_{i=1}^{B} \left(\hat{\theta}_j - \bar{\hat{\theta}}_j\right)^2}{B-1}} \quad (17)$$

where $\hat{\theta}_j$ is the parameter estimate obtained for a single replication, and $\bar{\hat{\theta}}_j$ is the mean parameter estimate obtained for *B* replications (Hwang, Malhotra, et al.,, 2010; Sharma, Durvasula, & Dillon, 1989; Srinivasan & Mason, 1986). MAD was calculated separately for the measurement and structural models in each replication data set.

The ability of the estimation methods to produce standard errors was also evaluated by constructing a confidence interval around each parameter estimate and determining whether the corresponding population parameter falls within this confidence interval (i.e., accuracy of the standard error estimate; Gerbing & Anderson, 1985). For this purpose, the confidence interval was defined as ± 1.96 standard errors around the parameter estimate, and the value of interest is the proportion of parameter estimates for which the population parameter falls within the appropriate confidence interval. This value was calculated for each replication to reflect the accuracy of the standard errors associated with the measurement and structural models separately.

Analytic Approach

To evaluate the performance of the four estimation methods, a multivariate analysis of variance (MANOVA) was calculated, and included the five independent variables of interest as factors (i.e., sample size, number of items per latent variable, degree of misspecification, type of latent variable-indicator relationships, and estimation method), and the seven key outcomes of interest described above as the dependent variables (i.e., GOF, average measurement model bias, average structural model bias, MAD of measurement model standard error estimates, MAD of structural model standard error estimates, accuracy of standard error estimates for the measurement model, and

accuracy of standard error estimates for the structural model). All interaction effects were included in the MANOVA. Effect sizes (partial η^2) were calculated for each direct and interaction effect. This method is consistent with recommendations and practices in this field (e.g., Hwang, Malhotra, et al., 2010; Paxton et al., 2001), and strengthens the connection between this and previous work.

CHAPTER IV. RESULTS

Analytic Procedure

A five-factor MANOVA was computed as a first step to understanding the effects of sample size, number of items per latent variable, degree of misspecification, type of latent variable-indicator relationships, and estimation method within the present study. The outcomes of interest included in the MANOVA were bias of GOF, bias in the parameter estimates of the measurement model, bias in the parameter estimates of the structural model, MAD of the standard error estimates associated with the measurement model, MAD of the standard error estimates associated with the structural model, accuracy of the estimates recovered for the measurement model, and accuracy of the estimates recovered for the structural model. The results of the multivariate tests are displayed in Table A.1.

It is important to note that the significant effects observed for this model may be merely a reflection of the large number of observations included in the complete data set for this study (a total of 21,600 observations representing 150 replications for each of 36 experimental design conditions for each of four estimation methods). For this reason, only significant results for which the tests of between-subjects effects were characterized by a medium or large effect size (i.e., partial $\eta^2 \geq .06$) will be presented and discussed (Hwang, Malhotra, et al., 2010; Paxton et al., 2001). In instances where pairwise comparisons are made, only significant results for which the differences are characterized by a medium or large effect size (i.e., $d \geq .50$) will be presented (Cohen, 1988). Accordingly, direct and interaction effects are described as moderate or large and not as

significant or not significant. Because interpretation of the results presented herein is based on the effect size of each effect/difference, no attempt was made to control for the overall family wise error rate associated with *p* values when multiple analyses are conducted. Where relevant, *p* values are reported as a matter of standard practice, and not for the purpose of interpreting or understanding effects.

Results by Outcome

Of the independent variables of interest in the present study (sample size, number of items per latent variable, degree of model misspecification, nature of latent variable-indicator relationships, estimation method), two consist of non-ordered categories (i.e., model misspecification and type of latent variable-indicator relationships). For the purposes of simplifying and organizing the presentation of results, effects are presented and discussed within the context of the four categories of models created by these two independent variables (i.e., correctly specified models with reflective indicators, correctly specified models with formative indicators, misspecified models with reflective indicators, misspecified models with formative indicators), except for the results related to model convergence.

Model Convergence

Model convergence represents the proportion of replications for which an estimation method was able to produce estimates of the model's parameters and standard errors within 1000 iterations, and those estimates were found to consist of statistically plausible values. All replications (total of 5,400 per estimation method) converged successfully for PLS and GSCA methods; 94% of ML and 87.7% of MCMC replications

converged successfully. Table 1 displays the number of replications in each experimental condition that converged to plausible values for ML and MCMC. This information is depicted graphically in Figure 6.

Table 1. Number of successfully converged replications by estimation method and experimental condition

		n = 50			*n* = 300			*n* = 1000		
		i = 3	*i* = 5	*i* = 7	*i* = 3	*i* = 5	*i* = 7	*i* = 3	*i* = 5	*i* = 7
ML	Correct Specification									
	Reflective	145	150	150	150	150	150	150	150	150
	Formative	150	146	132	150	147	148	150	149	145
	Misspecification									
	Reflective	141	148	102	150	150	150	150	150	150
	Formative	35	147	134	58	149	147	119	150	146
MCMC	Correct Specification									
	Reflective	147	97	51	150	150	150	150	150	150
	Formative	150	39	149	150	150	150	150	150	150
	Misspecification									
	Reflective	150	124	102	150	150	150	150	150	150
	Formative	149	33	147	132	149	78	150	150	72

For ML estimation, successful convergence was found to be influenced by number of items per latent variable ($F(2) = 368.44$, $p < .001$, partial $\eta^2 = 0.12$), degree of model misspecification ($F(1) = 418.68$, $p < .001$, partial $\eta^2 = 0.07$), and the nature of the latent variable-indicator relationships ($F(1) = 570.51$, $p < .001$, partial $\eta^2 = 0.10$). Generally stated, ML estimation converged for a larger proportion of replications as the number of items per latent variable increased. ML was also more likely to converge successfully when the model was correctly specified and when it included reflective instead of formative indicators. For MCMC estimation, successful convergence was found to be influenced by sample size ($F(2) = 722.52$, $p < .001$, partial $\eta^2 = 0.21$) and the number of items per latent variable ($F(2) = 326.52$, $p < .001$, partial $\eta^2 = 0.11$).

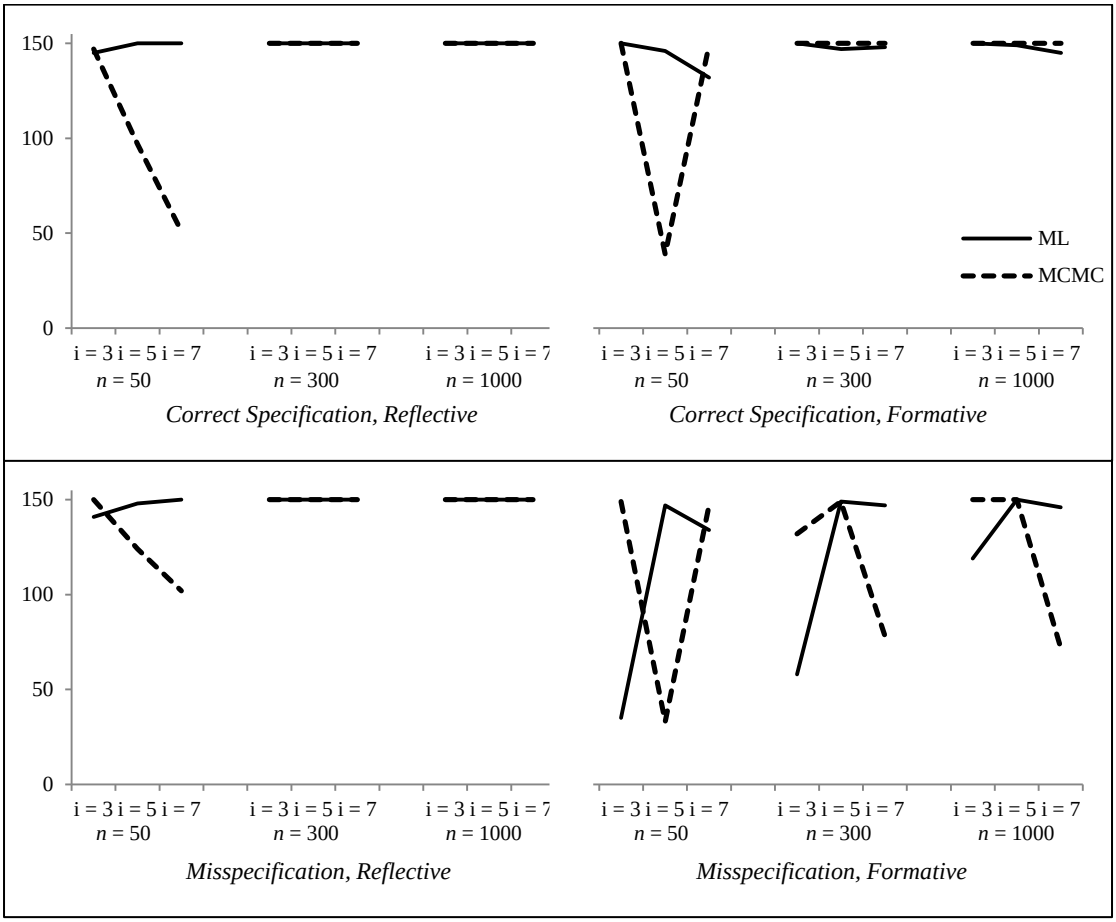


Figure 6. Number of successfully converged replications by condition.

Generally stated, the MCMC approach converged a higher proportion of times as both sample size and number of items per latent variable increased, except in situations when the sample size was very small (i.e., $n = 50$), and under conditions of model misspecification with formative indicators, as seen in Figure 6.

Goodness of Fit

Goodness of fit was evaluated in terms of the bias of the recovered estimate of model fit for each replication, by calculating the difference between the GOF estimate for each replication and its true value. GOF estimates smaller than the true GOF value for the population model are described as underestimated, or negatively biased estimates; GOF

estimates larger than the true GOF value for the population model are described as overestimated, or positively biased estimates. Thus, evaluation of GOF was completed by comparing the amount of bias in the GOF estimates produced by each estimation method across the different levels of the independent variables. Mean GOF bias for each experimental condition is displayed by estimation method in Table 2.

Table 2. Mean Goodness of Fit bias by estimation method and experimental condition

		n = 50			*n* = 300			*n* = 1000		
		i = 3	*i* = 5	*i* = 7	*i* = 3	*i* = 5	*i* = 7	*i* = 3	*i* = 5	*i* = 7
ML	Correct Specification									
	Reflective	-.183 (*SD* = 0.088)	-.181 (*SD* = 0.078)	-.172 (*SD* = 0.068)	-.181 (*SD* = 0.034)	-.182 (*SD* = 0.027)	-.175 (*SD* = 0.031)	-.180 (*SD* = 0.017)	-.180 (*SD* = 0.016)	-.177 (*SD* = 0.015)
	Formative	-.012 (*SD* = 0.078)	.002 (*SD* = 0.080)	.000 (*SD* = 0.071)	-.001 (*SD* = 0.030)	-.002 (*SD* = 0.026)	.001 (*SD* = 0.029)	.002 (*SD* = 0.009)	-.001 (*SD* = 0.015)	.002 (*SD* = 0.015)
	Misspecification									
	Reflective	.139 (*SD* = 0.071)	.118 (*SD* = 0.069)	.078 (*SD* = 0.065)	.146 (*SD* = 0.028)	.117 (*SD* = 0.023)	.079 (*SD* = 0.028)	.147 (*SD* = 0.014)	.119 (*SD* = 0.014)	.076 (*SD* = 0.013)
	Formative	.184 (*SD* = 0.055)	.059 (*SD* = 0.070)	.043 (*SD* = 0.075)	.182 (*SD* = 0.029)	.058 (*SD* = 0.028)	.044 (*SD* = 0.028)	.188 (*SD* = 0.014)	.058 (*SD* = 0.015)	.044 (*SD* = 0.015)
PLS	Correct Specification									
	Reflective	-.225 (*SD* = 0.072)	-.198 (*SD* = 0.069)	-.180 (*SD* = 0.061)	-.238 (*SD* = 0.031)	-.216 (*SD* = 0.026)	-.200 (*SD* = 0.029)	-.238 (*SD* = 0.017)	-.217 (*SD* = 0.016)	-.205 (*SD* = 0.014)
	Formative	-.012 (*SD* = 0.078)	.001 (*SD* = 0.079)	.001 (*SD* = 0.068)	-.002 (*SD* = 0.031)	-.002 (*SD* = 0.026)	.002 (*SD* = 0.029)	.000 (*SD* = 0.016)	-.001 (*SD* = 0.015)	.002 (*SD* = 0.015)
	Misspecification									
	Reflective	.104 (*SD* = 0.064)	.093 (*SD* = 0.064)	.063 (*SD* = 0.059)	.099 (*SD* = 0.027)	.081 (*SD* = 0.023)	.049 (*SD* = 0.027)	.100 (*SD* = 0.014)	.081 (*SD* = 0.014)	.044 (*SD* = 0.013)
	Formative	.262 (*SD* = 0.046)	.174 (*SD* = 0.050)	.128 (*SD* = 0.057)	.262 (*SD* = 0.018)	.173 (*SD* = 0.019)	.130 (*SD* = 0.022)	.264 (*SD* = 0.010)	.173 (*SD* = 0.010)	.130 (*SD* = 0.012)
GSCA	Correct Specification									
	Reflective	-.232 (*SD* = 0.078)	-.206 (*SD* = 0.075)	-.179 (*SD* = 0.085)	-.238 (*SD* = 0.032)	-.217 (*SD* = 0.026)	-.209 (*SD* = 0.043)	-.251 (*SD* = 0.002)	-.217 (*SD* = 0.016)	-.207 (*SD* = 0.019)
	Formative	-.004 (*SD* = 0.076)	.012 (*SD* = 0.080)	.010 (*SD* = 0.088)	.000 (*SD* = 0.030)	.000 (*SD* = 0.026)	-.147 (*SD* = 0.032)	.001 (*SD* = 0.016)	-.001 (*SD* = 0.015)	-.012 (*SD* = 0.019)
	Misspecification									
	Reflective	.101 (*SD* = 0.069)	.093 (*SD* = 0.068)	.053 (*SD* = 0.070)	.100 (*SD* = 0.028)	.082 (*SD* = 0.024)	.042 (*SD* = 0.040)	.100 (*SD* = 0.015)	.082 (*SD* = 0.014)	.037 (*SD* = 0.018)
	Formative	.313 (*SD* = 0.042)	.229 (*SD* = 0.047)	.161 (*SD* = 0.077)	.314 (*SD* = 0.016)	.228 (*SD* = 0.018)	.166 (*SD* = 0.027)	.315 (*SD* = 0.008)	.228 (*SD* = 0.010)	.166 (*SD* = 0.014)
MCMC	Correct Specification									
	Reflective	-.200 (*SD* = 0.083)	-.182 (*SD* = 0.077)	-.166 (*SD* = 0.061)	-.186 (*SD* = 0.034)	-.185 (*SD* = 0.027)	-.175 (*SD* = 0.031)	-.179 (*SD* = 0.017)	-.183 (*SD* = 0.016)	-.179 (*SD* = 0.014)
	Formative	-.015 (*SD* = 0.077)	.112 (*SD* = 0.049)	.000 (*SD* = 0.068)	-.001 (*SD* = 0.030)	-.002 (*SD* = 0.026)	.001 (*SD* = 0.029)	.001 (*SD* = 0.016)	-.001 (*SD* = 0.015)	-.002 (*SD* = 0.015)
	Misspecification									
	Reflective	.124 ***(SD = 0.072)***	.102 ***(SD = 0.068)***	.074 ***(SD = 0.059)***	.141 ***(SD = 0.028)***	.113 ***(SD = 0.023)***	.078 ***(SD = 0.028)***	.148 ***(SD = 0.014)***	.116 ***(SD = 0.014)***	.074 ***(SD = 0.013)***
	Formative	.211 (*SD* = 0.050)	.137 (*SD* = 0.098)	.039 (*SD* = 0.075)	.211 (*SD* = 0.023)	.058 (*SD* = 0.028)	.050 (*SD* = 0.029)	.209 (*SD* = 0.016)	.058 (*SD* = 0.015)	.041 (*SD* = 0.015)

In the overall MANOVA conducted for this study, no effect of estimation method on GOF bias was found ($F(3) = 68.73$, $p < .001$, partial $\eta^2 = 0.01$). However, a moderate effect was found for the degree of misspecification × estimation method interaction ($F(3) = 769.61$, $p < .001$, partial $\eta^2 = 0.10$), and a large effect was found for the latent variable-indicator relationship × estimation method interaction ($F(3) = 1296.56$, $p < .001$, partial $\eta^2 = 0.16$). The degree of misspecification × latent variable-indicator relationship × estimation method interaction was also found to be moderate ($F(3) = 718.83$, $p < .001$, partial $\eta^2 = 0.10$).

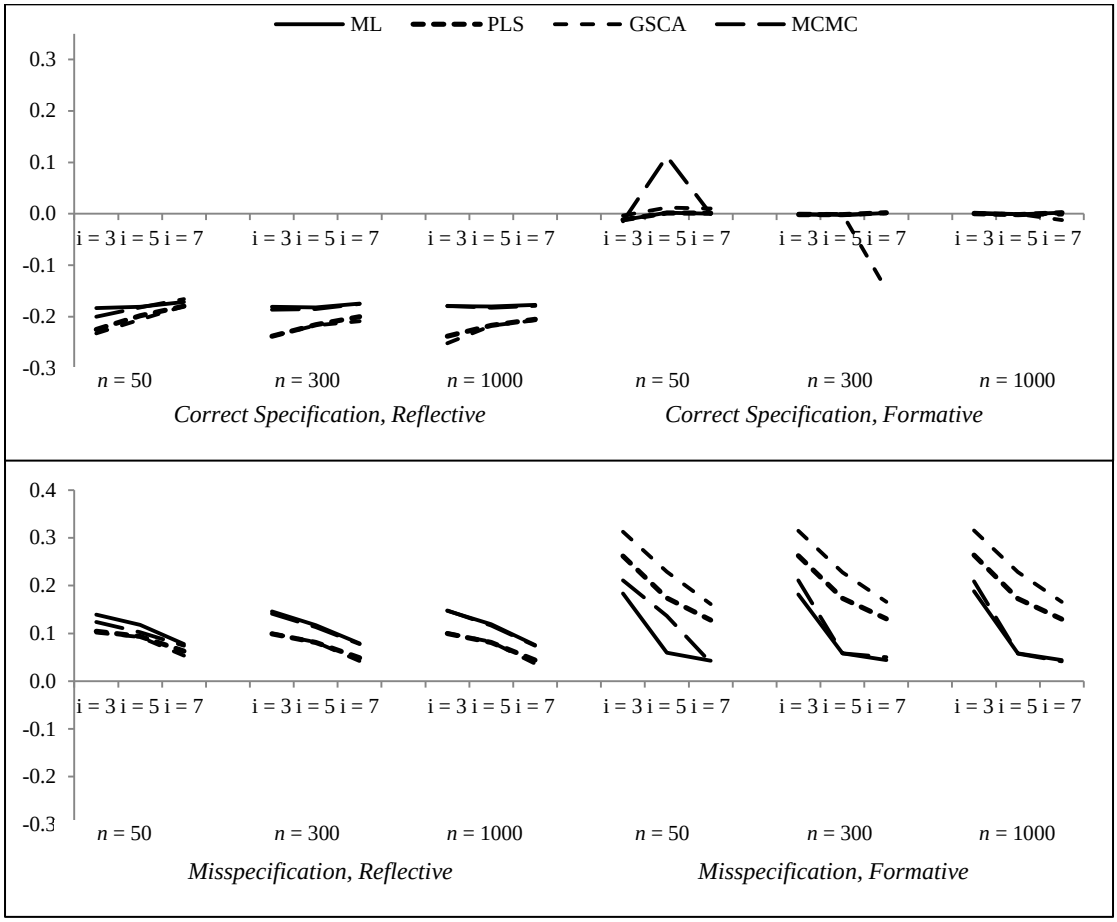


Figure 7. Bias of Goodness of Fit Estimates.

Figure 7 depicts the amount of bias observed for GOF estimates produced by each estimation method, by experimental condition.

Correct Specification, Reflective Indicators. Pair-wise post hoc comparisons indicated that GOF was consistently underestimated (i.e., yielded the most negative bias) under conditions of correct model specification and reflective measurement model relationships, regardless of sample size or number of items. GOF bias for correctly specified models with reflective indicators was found to be different from correctly specified models with formative indicators (mean difference = -0.194, $p < .001$, $d = 3.67$), misspecified models with reflective indicators (mean difference = -0.293, $p < .001$, $d = 5.66$), and misspecified models with formative indicators (mean difference = -0.359, $p < .001$, $d = 4.67$) regardless of sample size or number of items per latent variable, with GOF for correctly specified reflective models underestimated across all levels of sample size and number of items per latent variable. Differences in the bias of GOF estimates for correctly specified reflective models were identified between estimation methods. Under these conditions, PLS and GSCA produced more biased (i.e., more underestimated) estimates of GOF than ML and MCMC across all levels of sample size and number of indicators ($F(1) = 714.51$, $p < .001$, partial $\eta^2 = 0.12$). Bias in the GOF estimates was found to decrease for PLS and GSCA as number of items increased, ($F(2) = 169.48$, $p < .001$, partial $\eta^2 = 0.11$).

Correct Specification, Formative Indicators. Descriptively, GOF bias was smallest under conditions of correct model specification with formative measurement models ($M = -0.0049$, $SD = 0.05$) across all levels of sample size, number of items per

latent variable, and estimation method. Within this set of conditions, ML and PLS recovered approximately equal estimates of GOF across all levels of sample size and number of indicators, with the average bias for GOF estimates close to zero for each of these two methods. MCMC recovered GOF estimates similar to ML and PLS across levels of number of items, except for when sample size was smallest, as seen in Figure 7. A large sample size × number of indicators per latent variable interaction effect was found for GSCA ($F(4) = 135.06$, $p < .001$, partial $\eta^2 = 0.29$), which produced more biased (i.e., more underestimated) estimates of GOF for seven items compared to three or five items when sample size was larger than 50. No change in bias was observed across number of indicators per latent variable with sample size was smallest (i.e., $n = 50$). This indicates that the higher level of bias observed for the MCMC method when $i = 5$ is not meaningfully different from the bias observed for the other two levels of i (i.e., 3, 7). A moderate effect of number of items was found for MCMC ($F(2) = 36.16$, $p < .001$, partial $\eta^2 = 0.06$).

Misspecification, Reflective Indicators. A follow-up univariate ANOVA indicated that ML performed similarly to MCMC, and PLS performed similarly to GSCA when estimating GOF for misspecified models. For further consideration of GOF for misspecified models, the four estimation methods were combined into two levels of a single predictor (i.e., ML and MCMC were combined, PLS and GSCA were combined), and a univariate ANOVA was calculated to evaluate the effects of sample size and number of items on GOF bias. Under conditions of misspecified models and reflective indicators, bias of GOF estimates produced by ML and MCMC was found to decrease as

the number of items increased ($F(2) = 514.23$, $p < .001$, partial $\eta^2 = 0.28$), but sample size was not found to have any effect on GOF estimate bias (partial $\eta^2 < 0.06$). A similar pattern of results was found for PLS and GSCA, where bias was found to decrease as the number of items increased ($F(2) = 370.77$, $p < .001$, partial $\eta^2 = 0.21$), but sample size was not found to have any effect on GOF estimate bias (partial $\eta^2 < 0.06$). Despite the similarity in the patterns of results between the two sets of estimation methods, a moderate simple effect indicated that ML and MCMC performed differently from PLS and GSCA ($F(1) = 795.12$, $p < .001$, partial $\eta^2 = 0.13$), with PLS and GSCA consistently recovering less biased parameter estimates than ML and MCMC.

Misspecification, Formative Indicators. Under conditions of misspecified models with formative indicators, bias of the GOF estimates produced by ML and MCMC was found to decrease as the number of items increased ($F(2) = 2982.71$, $p < .001$, partial $\eta^2 = 0.74$). A similar pattern of results was found for PLS and GSCA, where bias was found to decrease as the number of items increased ($F(2)$ 2640.16, $p < .001$, partial $\eta^2 = 0.66$), with less bias for 5 items compared to 3 items ($p < .001$, $d = 2.22$), and even less bias for 7 items compared to 5 items ($p < .001$, $d = 1.24$). Despite the similarity in the patterns of results between the two sets of estimation methods, a large effect of estimation method group indicated that ML and MCMC performed differently from PLS and GSCA ($F(1) = 8151.47$, $p < .001$, partial $\eta^2 = 0.63$), with ML and MCMC consistently recovering less biased parameter estimates than PLS and GSCA.

Bias of Measurement Model Parameter Estimates

Recovery of measurement model parameters was evaluated in terms of the relative bias of the parameter estimates, given the true values of the parameters (refer to equation 15). In the overall MANOVA conducted for this study, the simple effect of estimation method on measurement model bias was found to be large ($F(3) = 20046.30$, $p < .001$, partial $\eta^2 = 0.75$). Moderate and large interactions were also identified between estimation method and sample size ($F(6) = 1040.21$, $p < .001$, partial $\eta^2 = 0.23$), number of items per latent variable ($F(6) = 3181.93$, $p < .001$, partial $\eta^2 = 0.48$), degree of model misspecification ($F(3) = 434.75$, $p < .001$, partial $\eta^2 = 0.06$), and nature of the latent variable-indicator ($F(3) = 2507.73$, $p < .001$, partial $\eta^2 = 0.27$). Bias of the measurement model parameter estimates across all levels of the independent variables are displayed in Table 3 and depicted in Figure 8.

Follow-up analyses indicated no differences in bias of the measurement model parameters between the ML or MCMC approaches for correctly specified models with reflective indicators ($F(1) = 15.57$, partial $\eta^2 < 0.06$), correctly specified models with formative indicators ($F(1) = 0.91$, partial $\eta^2 < 0.06$), misspecified models with reflective indicators ($F(1) = 1.85$, partial $\eta^2 < 0.06$), or misspecified models with formative indicators ($F(1) = 6.83$, partial $\eta^2 < 0.06$). Thus, results of the ML and MCMC estimation methods were combined for further exploration of the performance of the estimation methods in the recovery of measurement model parameters.

Table 3. Mean bias of measurement model parameter estimates by estimation method and experimental condition

		n = 50			*n* = 300			*n* = 1000		
		i = 3	*i* = 5	*i* = 7	*i* = 3	*i* = 5	*i* = 7	*i* = 3	*i* = 5	*i* = 7
ML	Correct Specification									
	Reflective	12.096 (*SD* = 3.696)	10.442 (*SD* = 2.654)	9.847 (*SD* = 1.881)	4.784 (*SD* = 1.443)	4.091 (*SD* = 0.883)	3.968 (*SD* = 0.743)	2.712 (*SD* = 0.719)	2.261 (*SD* = 0.556)	2.117 (*SD* = 0.408)
	Formative	.108 (*SD* = 0.017)	.109 (*SD* = 0.014)	.106 (*SD* = 0.012)	.100 (*SD* = 0.002)	.100 (*SD* = 0.001)	.100 (*SD* =0 .001)	.100 (*SD* = 0.000)	.100 (*SD* = 0.000)	.100 (*SD* =0 .000)
	Misspecification									
	Reflective	14.320 (*SD* = 2.883)	12.140 (*SD* = 2.496)	11.207 (*SD* = 1.765)	10.758 (*SD* = 1.439)	7.486 (*SD* = 0.914)	6.264 (*SD* = 0.686)	9.673 (*SD* = 0.871)	6.103 (*SD* = 0.562)	4.741 (*SD* = 0.419)
	Formative	21.558 (*SD* = 0.988)	11.959 (*SD* = 0.745)	8.589 (*SD* = 0.589)	21.522 (*SD* = 0.512)	12.001 (*SD* = 0.311)	8.634 (*SD* = 0.227)	21.643 (*SD* = 0.237)	12.003 (*SD* = 0.169)	8.625 (*SD* = 0.123)
PLS	Correct Specification									
	Reflective	16.584 (*SD* = 2.885)	11.670 (*SD* = 2.122)	10.300 (*SD* = 1.705)	15.704 (*SD* = 1.249)	9.533 (*SD* = 1.248)	7.259 (*SD* = 1.343)	15.896 (*SD* = 0.737)	9.715 (*SD* = 0.771)	7.064 (*SD* = 0.784)
	Formative	.070 (*SD* = 0.012)	.084 (*SD* = 0.013)	.088 (*SD* = 0.014)	.069 (*SD* = 0.005)	.083 (*SD* = 0.005)	.088 (*SD* = 0.006)	.068 (*SD* = 0.002)	.083 (*SD* = 0.003)	.088 (*SD* = 0.003)
	Misspecification									
	Reflective	19.501 (*SD* = 2.445)	13.805 (*SD* = 2.112)	11.925 (*SD* = 1.708)	19.043 (*SD* = 1.170)	12.044 (*SD* = 1.212)	9.249 (*SD* = 1.326)	19.196 (*SD* = 0.689)	12.234 (*SD* = 0.759)	9.040 (*SD* = 0.769)
	Formative	23.867 (*SD* = 0.576)	14.895 (*SD* = 0.220)	11.518 (*SD* = 0.166)	23.853 (*SD* = 0.221)	14.881 (*SD* = 0.091)	11.500 (*SD* = 0.069)	23.860 (*SD* = 0.107)	14.877 (*SD* = 0.053)	11.503 (*SD* = 0.036)
GSCA	Correct Specification									
	Reflective	16.084 (*SD* = 2.464)	11.079 (*SD* = 2.044)	17.420 (*SD* = 2.591)	15.804 (*SD* = 1.261)	9.498 (*SD* = 1.279)	12.236 (*SD* = 1.651)	15.918 (*SD* = 0.731)	9.734 (*SD* = 0.766)	13.667 (*SD* = 1.090)
	Formative	.115 (*SD* = 0.031)	.126 (*SD* = 0.023)	5.033 (*SD* = 1.094)	.081 (*SD* = 0.011)	.098 (*SD* = 0.007)	27.112 (*SD* = 0.802)	.087 (*SD* = 0.009)	.100 (*SD* = 0.001)	4.848 (*SD* = 0.218)
	Misspecification									
	Reflective	19.262 (*SD* = 2.415)	13.456 (*SD* = 2.073)	11.724 (*SD* = 1.105)	19.078 (*SD* = 1.164)	12.015 (*SD* = 1.230)	14.139 (*SD* = 1.680)	19.196 (*SD* = 0.671)	12.207 (*SD* = 0.752)	13.667 (*SD* = 1.090)
	Formative	25.320 (*SD* = 0.496)	14.971 (*SD* = 0.358)	15.255 (*SD* = 1.245)	25.359 (*SD* = 0.178)	14.985 (*SD* = 0.130)	15.243 (*SD* = 0.494)	25.362 (*SD* = 0.086)	14.999 (*SD* = 0.068)	15.246 (*SD* = 0.281)
MCMC	Correct Specification									
	Reflective	12.588 (*SD* = 3.750)	11.789 (*SD* = 3.372)	10.202 (*SD* = 1.977)	4.916 (*SD* = 1.513)	4.198 (*SD* = 0.945)	4.037 (*SD* = 0.736)	2.738 (*SD* = 0.706)	2.313 (*SD* = 0.561)	2.136 (*SD* = 0.409)
	Formative	.110 (*SD* = 0.018)	.109 (*SD* = 0.012)	.106 (*SD* = 0.013)	.100 (*SD* = 0.002)	.100 (*SD* = 0.001)	.100 (*SD* = 0.001)	.100 (*SD* = 0.000)	.100 (*SD* = 0.000)	.100 (*SD* = 0.000)
	Misspecification									
	Reflective	14.287 (*SD* = 3.148)	12.848 (*SD* = 3.070)	11.205 (*SD* = 2.056)	10.794 (*SD* = 1.469)	7.509 (*SD* = 0.933)	6.305 (*SD* = 0.705)	9.637 (*SD* = 0.855)	6.151 (*SD* = 0.573)	4.745 (*SD* = 0.426)
	Formative	21.225 (*SD* = 1.346)	12.592 (*SD* = 0.529)	8.497 (*SD* = 0.618)	21.238 (*SD* = 0.502)	12.009 (*SD* = 0.310)	8.648 (*SD* = 0.246)	21.378 (*SD* = 0.282)	11.787 (*SD* = 0.597)	7.920 (*SD* = 0.843)

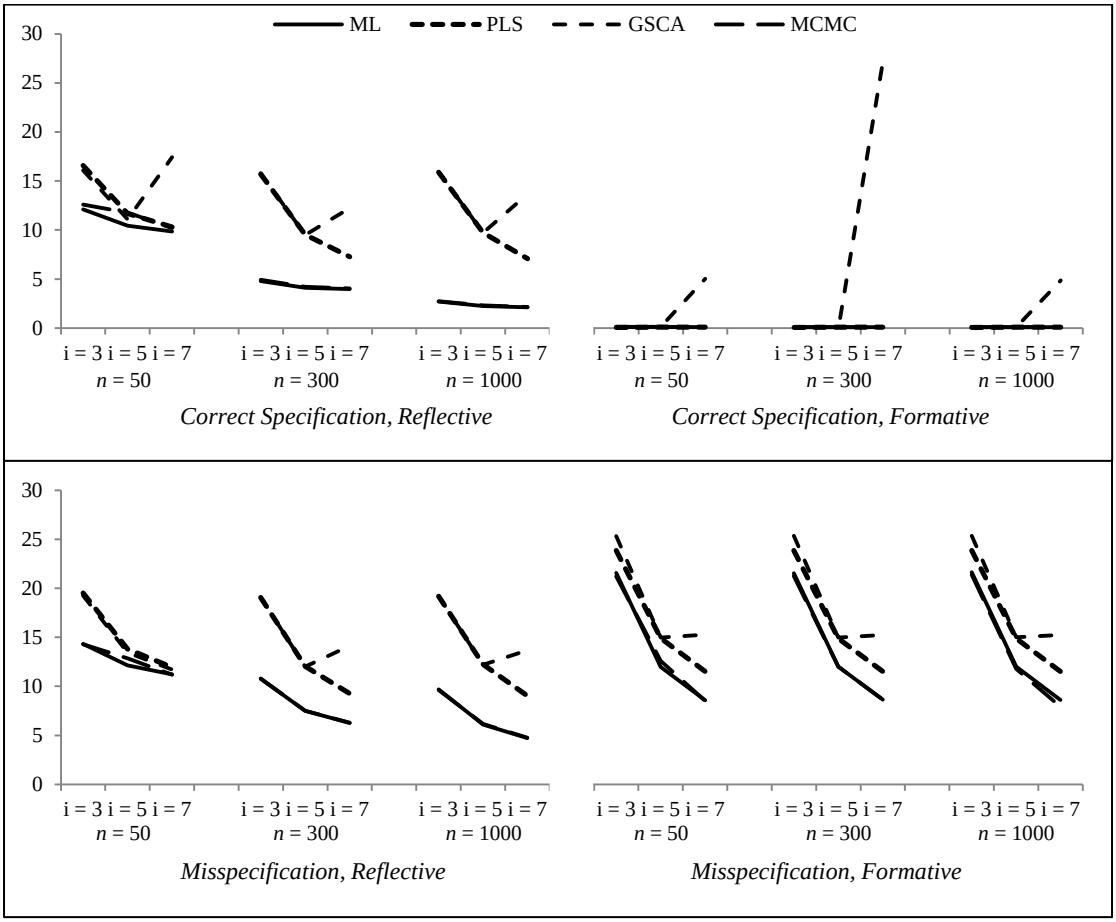


Figure 8. Bias of Measurement Model Parameter Estimates.

Correct Specification, Reflective Indicators. Within the context of correctly specified models with reflective relationships in the measurement model, bias in the parameter estimates for the measurement model was found to decrease as sample size increased ($F(2) = 2749.55$, $p < .001$, partial $\eta^2 = 0.51$) and number of items increased ($F(2) = 2712.43$, $p < .001$, partial $\eta^2 = 0.51$). For ML/MCMC, a large decrease in measurement model bias was found when sample size increased from $n = 50$ to $n = 300$ ($d = 2.85$), and from $n = 300$ to $n = 1000$ ($d = 2.11$). For PLS, the increase in sample size from $n = 50$ to $n = 300$ was found to be moderate ($d = 0.55$). A moderate decrease in measurement model bias was observed for GSCA as the sample size increased from $n =$

50 to n = 300 (d = 0.71). Large decreases in bias were also observed for ML/MCMC and PLS as number of items increased from i = 3 to i = 5 (ML/MCMC: d = 0.25; PLS: d = 3.13) and from i = 5 to i = 7 (ML/MCMC: d = 0.16; PLS: d = 1.11). For GSCA, a large decrease in measurement model bias was observed as the number of items increased from i = 3 to i = 5 (d = 3.57), but this was followed by a significant increase in bias as the number of items continued to increase from i = 5 to i = 7 (d = 1.86). Across all levels of sample size and number of indicators, ML/MCMC produced less biased estimates than either PLS (d = 0.47) or GSCA (d = 2.09). Across all estimation methods and levels of sample size and number of indicators, bias of the measurement model parameter estimates was positive, which indicates that ML, PLS, GSCA, and MCMC consistently overestimated model parameters regardless of sample size or number of indicators per item.

Correct Specification, Formative Indicators. Within the context of correctly specified models with formative relationships in the measurement model, sample size ($F(2) = 59572.79$, $p < .001$, partial $\eta^2 = 0.96$) and number of items ($F(2) = 163608.88$, $p < .001$, partial $\eta^2 = 0.98$) were found to effect bias of the parameter estimates for the measurement model. For ML/MCMC, measurement model bias was found to decrease as sample size increased ($F(2) = 234.98$, $p < .001$, partial $\eta^2 = 0.16$). For PLS, measurement model bias was found to increase as the number of items increased ($F(2) = 634.15$, $p < .001$, partial $\eta^2 = 0.49$). For GSCA, sample size and number of items were found to have large effects on measurement model bias ($F(2) = 38896.91$, $p < .001$, partial $\eta^2 = 0.98$ and $F(2) = 106838.42$, $p < .001$, partial $\eta^2 = 0.99$, respectively).

Despite these differences between estimation methods and differences in method recovery of measurement model parameters across levels of sample size and number of indicators per latent variable, ML, PLS, and MCMC recovered parameter estimates with almost no bias for the measurement model across all levels of sample size and number of items GSCA also recovered parameter estimates for the measurement model with close to no bias when the number of items per latent variable were small, but overestimated measurement model parameters for a larger number of items (*i.e.,* 7) across all sample sizes.

Misspecification, Reflective Indicators. Within the context of misspecified models with reflective relationships in the measurement model, large main effects were identified for sample size ($F(2) = 1069.52$, $p < .001$, partial $\eta^2 = 0.29$), number of items ($F(2) = 7907.84$, $p < .001$, partial $\eta^2 = 0.75$), and estimation method ($F(2) = 7691.74$, $p < .001$, partial $\eta^2 = 0.74$). In addition, the interaction between estimation method and sample size was found to be large ($F(4) = 687.12$, $p < .001$, partial $\eta^2 = 0.34$), as was the interaction between estimation method and number of items ($F(4) = 548.98$, $p < .001$, partial $\eta^2 = 0.29$). For ML/MCMC estimation, a significant effect of sample size was identified ($F(2) = 2890.62$, $p < .001$, partial $\eta^2 = 0.69$), whereby bias in the parameter estimates for the measurement model were found to decrease as sample size increased. A large effect of number of items was also identified for ML/MCMC estimation ($F(2) = 1472.71$, $p < .001$, partial $\eta^2 = 0.53$), where measurement model bias was found to decrease as number of items increased, regardless of sample size. For PLS estimation, large effects were also identified for sample size ($F(2) = 178.59$, $p < .001$, partial $\eta^2 =$

0.21) and number of items ($F(2) = 4616.75$, $p < .001$, partial $\eta^2 = 0.87$), with a decrease in measurement model parameter estimate bias observed as sample size and number of indicators increase. Further, a moderate sample size × number of items interaction effect was identified for PLS ($F(4) = 33.94$, $p < .001$, partial $\eta^2 = 0.09$). Bias in the parameter estimates for the measurement model recovered by PLS was found to decrease as both sample size and number of items increased, with the rate of decrease over increased number of items being more severe as sample size increased. For GSCA parameter recovery, a large effect of number of items on measurement model parameter estimate bias was identified ($F(2) = 2813.73$, $p < .001$, partial $\eta^2 = 0.81$), as was a large sample size × number of items interaction effect ($F(4) = 77.21$, $p < .001$, partial $\eta^2 = 0.19$). Across all levels of sample size, bias in GSCA parameter estimates decreased as number of items increased from $i = 3$ to $i = 5$, but bias only continued to decrease as number of items increased from $i = 5$ to $i = 7$ when $n = 50$. When a larger sample size was used (i.e., $n = 300$ or 1000), bias in the measurement model parameter estimates increased as number of items increased from $i = 5$ to $i = 7$.

Misspecification, Formative Indicators. Within the context of misspecified models with formative relationships in the measurement model, large effects were identified for number of items ($F(2) = 242454.68$, $p < .001$, partial $\eta^2 = 0.99$), estimation method ($F(2) = 36783.25$, $p < .001$, partial $\eta^2 = 0.94$), and the interaction between number of items and estimation method ($F(4) = 2426.59$, $p < .001$, partial $\eta^2 = 0.67$). A large effect of number of items on measurement model bias was found for both the ML/MCMC ($F(2)$ 81054.61, $p < .001$, partial $\eta^2 = 0.99$) and PLS ($F(2) = 339940.02$, $p <$

.001, partial $\eta^2 = 1.00$) methods. For ML, PLS, and MCMC, measurement model estimates became less biased as the number of items increased, though ML/MCMC parameter estimates were consistently less biased than those recovered by PLS. For GSCA, a large effect of number of items on measurement model bias was identified ($F(2) = 61233.69$, $p < .001$, partial $\eta^2 = 0.99$), indicating a decrease in bias as the number of items increased. Thus, GSCA was found to overestimate measurement model parameters less for models with more indicators per latent variable.

Bias of Structural Model Parameter Estimates

Recovery of structural model parameters was evaluated in terms of the relative bias of the parameter estimates, given the true values of the parameters (refer to equation 15). In the overall MANOVA conducted for this study, the simple effect of estimation method on structural model bias was found to be moderate ($F(3) = 1051.43$, $p < .001$, partial $\eta^2 = 0.13$). A moderate interaction was identified between estimation method and nature of the latent variable-indicator ($F(3) = 540.98$, $p < .001$, partial $\eta^2 = 0.07$). The amount of structural model parameter estimate bias produced by each estimation method, by experimental condition is depicted in Figure 9 and provided in Table 4.

Follow-up analyses indicated no differences in bias of the structural model parameters between the ML and MCMC approaches for correctly specified models with reflective indicators ($F(1) = 0.39$, partial $\eta^2 < 0.06$), correctly specified models with formative indicators ($F(1) = 12.66$, partial $\eta^2 < 0.06$), or misspecified models with reflective indicators ($F(1) = 2.37$, partial $\eta^2 < 0.06$). Similarly, no differences in bias of the structural model parameters were found between the PLS and GSCA approaches for

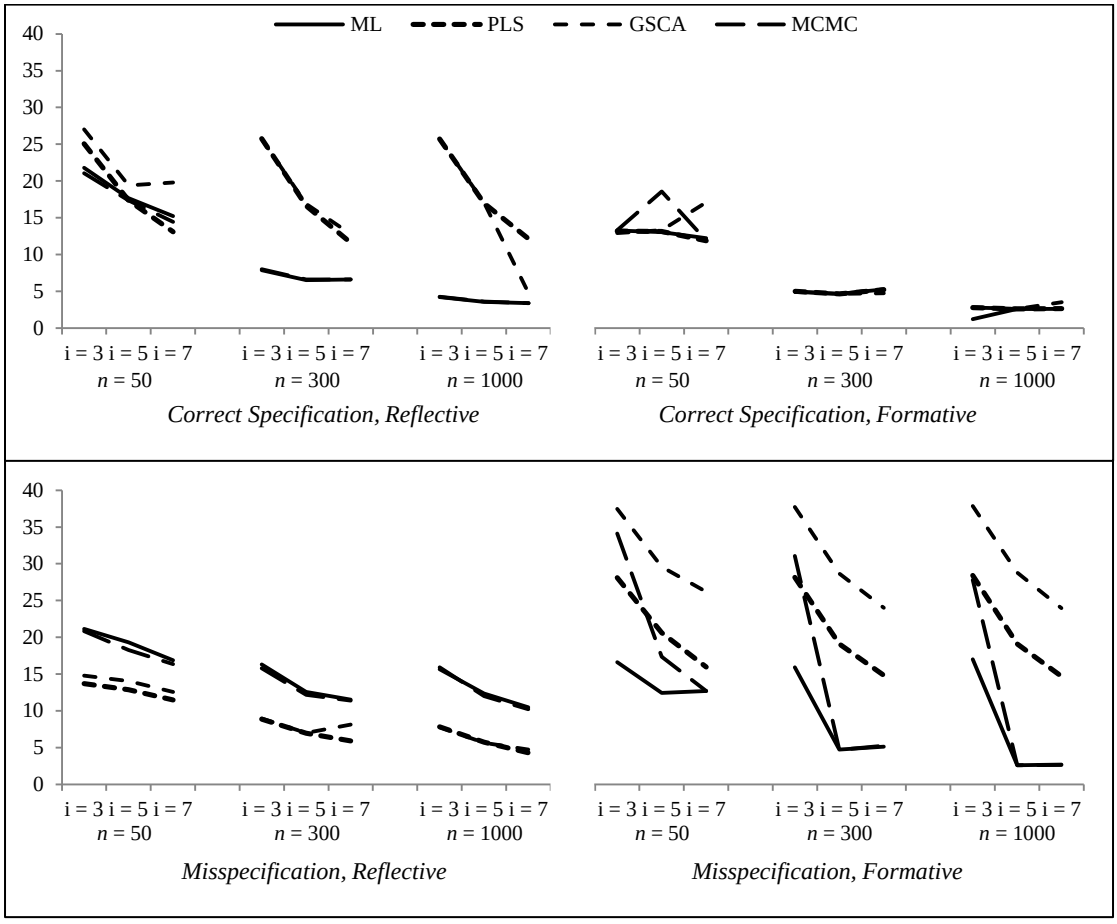


Figure 9. Bias of Structural Model Parameter Estimates.

correctly specified models with reflective indicators ($F(1) = 3.09$, partial $\eta^2 < 0.06$), correctly specified models with formative indicators ($F(1) = 10.70$, partial $\eta^2 < 0.06$), or misspecified models with reflective indicators ($F(1) = 2.50$, partial $\eta^2 < 0.06$). Thus, results of the ML and MCMC estimation methods were combined, and the PLS and GSCA results were combined for further exploration of the performance of the estimation methods in the recovery of structural model parameters under conditions of correct model specification and model misspecification with reflective measurement model relationships.

Table 4. Mean bias of structural model parameter estimates by estimation method and experimental condition

		$n = 50$			$n = 300$			$n = 1000$		
		$i = 3$	$i = 5$	$i = 7$	$i = 3$	$i = 5$	$i = 7$	$i = 3$	$i = 5$	$i = 7$
ML	Correct Specification									
	Reflective	21.816 (*SD* = 12.614)	17.663 (*SD* = 9.754)	15.211 (*SD* = 9.007)	7.864 (*SD* = 4.424)	6.501 (*SD* = 3.431)	6.609 (*SD* = 3.458)	4.203 (*SD* = 2.253)	3.569 (*SD* = 1.896)	3.388 (*SD* = 1.731)
	Formative	13.247 (*SD* = 7.586)	13.026 (*SD* = 7.884)	12.210 (*SD* = 6.378)	4.943 (*SD* = 2.684)	4.658 (*SD* = 2.585)	5.252 (*SD* = 2.423)	1.196 (*SD* = 1.271)	2.623 (*SD* = 1.461)	2.623 (*SD* = 1.423)
	Misspecification									
	Reflective	21.125 (*SD* = 9.703)	19.296 (*SD* = 9.353)	16.897 (*SD* = 8.175)	16.312 (*SD* = 4.514)	12.549 (*SD* = 4.202)	11.541 (*SD* = 4.739)	15.652 (*SD* = 2.707)	12.334 (*SD* = 2.606)	10.466 (*SD* = 2.411)
	Formative	16.634 (*SD* = 7.668)	12.431 (*SD* = 6.905)	12.696 (*SD* = 6.280)	15.941 (*SD* = 4.571)	4.721 (*SD* = 2.525)	5.120 (*SD* = 2.444)	17.002 (*SD* = 2.303)	2.599 (*SD* = 1.356)	2.698 (*SD* = 1.393)
PLS	Correct Specification									
	Reflective	25.047 (*SD* = 12.950)	17.373 (*SD* = 9.425)	13.109 (*SD* = 8.148)	25.759 (*SD* = 5.978)	16.603 (*SD* = 5.116)	11.620 (*SD* = 5.284)	25.728 (*SD* = 3.180)	16.954 (*SD* = 3.060)	12.153 (*SD* = 2.783)
	Formative	13.199 (*SD* = 7.597)	13.137 (*SD* = 8.366)	11.877 (*SD* = 6.247)	5.007 (*SD* = 2.808)	4.651 (*SD* = 2.582)	5.218 (*SD* = 2.386)	2.768 (*SD* = 1.502)	2.615 (*SD* = 1.448)	2.652 (*SD* = 1.403)
	Misspecification									
	Reflective	13.709 (*SD* = 8.069)	12.872 (*SD* = 7.182)	11.472 (*SD* = 6.524)	8.857 (*SD* = 3.946)	6.936 (*SD* = 3.330)	5.915 (*SD* = 3.092)	7.823 (*SD* = 2.060)	5.738 (*SD* = 1.974)	4.283 (*SD* = 1.922)
	Formative	28.122 (*SD* = 7.064)	20.615 (*SD* = 6.792)	15.975 (*SD* = 7.459)	28.163 (*SD* = 2.884)	19.095 (*SD* = 3.059)	14.839 (*SD* = 3.609)	28.392 (*SD* = 1.554)	19.075 (*SD* = 1.689)	14.697 (*SD* = 1.926)
GSCA	Correct Specification									
	Reflective	27.030 (*SD* = 14.605)	19.387 (*SD* = 11.108)	19.809 (*SD* = 12.473)	25.715 (*SD* = 6.145)	16.858 (*SD* = 5.202)	12.786 (*SD* = 7.607)	25.701 (*SD* = 3.070)	17.032 (*SD* = 3.072)	4.747 (*SD* = 2.505)
	Formative	12.912 (*SD* = 7.333)	13.309 (*SD* = 8.305)	17.099 (*SD* = 8.745)	4.930 (*SD* = 2.666)	4.646 (*SD* = 2.560)	4.747 (*SD* = 2.505)	2.799 (*SD* = 1.498)	2.624 (*SD* = 1.450)	3.534 (*SD* = 1.882)
	Misspecification									
	Reflective	14.768 (*SD* = 9.062)	14.047 (*SD* = 8.051)	12.577 (*SD* = 3.736)	8.951 (*SD* = 4.047)	6.997 (*SD* = 3.429)	8.134 (*SD* = 4.030)	7.791 (*SD* = 2.116)	5.631 (*SD* = 2.014)	4.747 (*SD* = 2.505)
	Formative	37.472 (*SD* = 6.543)	29.585 (*SD* = 6.927)	26.190 (*SD* = 9.457)	37.731 (*SD* = 2.553)	28.686 (*SD* = 2.977)	24.019 (*SD* = 4.623)	37.851 (*SD* = 1.362)	28.793 (*SD* = 1.600)	23.991 (*SD* = 2.427)
MCMC	Correct Specification									
	Reflective	21.076 (*SD* = 12.467)	17.365 (*SD* = 9.885)	14.449 (*SD* = 8.766)	8.009 (*SD* = 4.604)	6.621 (*SD* = 3.536)	6.602 (*SD* = 3.492)	4.272 (*SD* = 2.246)	3.609 (*SD* = 1.976)	3.374 (*SD* = 1.776)
	Formative	13.287 (*SD* = 7.532)	18.576 (*SD* = 8.104)	11.896 (*SD* = 6.220)	4.945 (*SD* = 2.690)	4.658 (*SD* = 2.569)	5.267 (*SD* = 2.417)	2.779 (*SD* = 1.515)	2.629 (*SD* = 1.457)	2.607 (*SD* = 1.366)
	Misspecification									
	Reflective	20.849 (*SD* = 9.672)	18.289 (*SD* = 8.620)	16.339 (*SD* = 8.004)	15.783 (*SD* = 4.319)	12.182 (*SD* = 4.112)	11.397 (*SD* = 4.681)	15.927 (*SD* = 2.728)	11.996 (*SD* = 2.645)	10.221 (*SD* = 2.380)
	Formative	34.110 (*SD* = 7.970)	17.375 (*SD* = 9.723)	12.705 (*SD* = 6.080)	31.058 (*SD* = 4.500)	4.712 (*SD* = 2.521)	5.248 (*SD* = 2.518)	27.806 (*SD* = 6.039)	2.612 (*SD* = 1.372)	2.642 (*SD* = 1.476)

Correct Specification, Reflective Indicators. Under conditions of correct model specification and a reflective measurement model, a large effect was

observed for estimation method ($F(2) = 1032.85$, $p < .001$, partial $\eta^2 = 0.28$). A large decrease in the bias of recovered structural model estimates was observed for ML/MCMC methods as sample size increased ($F(2) = 1092.10$, $p < .001$, partial $\eta^2 = 0.46$). For PLS and GSCA methods, bias in the structural model estimates decreased as number of items increased ($F(2) = 662.92$, $p < .001$, partial $\eta^2 = 0.33$). ML, PLS, GSCA, and MCMC consistently overestimated parameter estimates across all levels of sample size and number of indicators per latent variable for correctly specified models with reflective measurement model relationships. Across all levels of sample size and number of indicators per latent variable, ML and MCMC recovered parameter estimates for the structural model with less bias than those recovered by either PLS ($d = 1.02$) or GSCA ($d = 0.97$).

Correct Specification, Formative Indicators. Under conditions of correct model specification and formative measurement models, a difference was not found between ML/MCMC and PLS/GSCA estimation methods ($F(1) = 6.90$, partial $\eta^2 = 0.00$) for the amount of bias in recovered parameter estimates for the structural model. A large effect of sample size was observed for ML ($F(2) = 651.52$, $p < .001$, partial $\eta^2 = 0.50$), PLS ($F(2) = 582.83$, $p < .001$, partial $\eta^2 = 0.47$), GSCA ($F(2) = 582.83$, $p < .001$, partial $\eta^2 = 0.47$), and MCMC ($F(2) = 677.32$, $p < .001$, partial $\eta^2 = 0.50$), which indicated a decrease in structural model bias as sample size increases for all four estimation methods as sample size increases. It is worth noting that MCMC structural model parameter estimates were more biased for $i = 5$ when $n = 50$ than either of the other number of items conditions within this sample size. The difference in bias, however, is relatively small.

Misspecification, Reflective Indicators. Under conditions of model misspecification and reflective indicators, large and moderate effects of sample size ($F(2) = 276.78$, $p < .001$, partial $\eta^2 = 0.18$) and number of items ($F(2) = 150.46$, $p < .001$, partial $\eta^2 = 0.10$) were identified for the ML/MCMC approaches. For both ML and MCMC, bias in the structural model parameter estimates was found to decrease as sample increased, as well as when number of items increased. A large effect of sample size on structural model bias was found for the PLS/GSCA approaches as well ($F(2) = 542.76$, $p < .001$, partial $\eta^2 = 0.29$), with bias decreasing as sample size increased. Although the same trend was observed for both pairs of estimation methods, the effect of method was found to be large ($F(1) = 1649.43$, $p < .001$, partial $\eta^2 = 0.24$), which indicates that a greater decrease in bias was associated with increased sample size for the PLS/GSCA methods. Across all levels of sample size and number of indicators per latent variable, PLS and GSCA recovered less biased parameter estimates for structural models compared to ML and MCMC, with PLS recovering slightly less biased estimates than GSCA.

Misspecification, Formative Indicators. Under conditions of model misspecification and formative latent variable-indicator relationships, large effects were found for number of items ($F(2) = 3402.26$, $p < .001$, partial $\eta^2 = 0.59$) and estimation method ($F(3) = 3214.76$, $p < .001$, partial $\eta^2 = 0.67$) on bias in recovered structural model parameter estimates. Across all levels of sample size and number of indicators per latent variable, all four estimation methods overestimated parameter estimates for the structural model, but ML and MCMC produced less biased parameter estimates for the structural

model parameters. A large interaction effect between number of items and estimation method was also observed ($F(6) = 180.89$, $p < .001$, partial $\eta^2 = 0.19$). For all estimation methods, a decrease in bias of structural model estimates was observed as number of items increased from i =3 to i = 5 (ML: d = 1.98; PLS: d = 1.93; GSCA, d = 2.02; MCMC d = 5.25). For PLS and GSCA, bias continued to decrease as number of items increased from i =5 to i = 7 (PLS: d = 0.94; GSCA, d = 0.79); for MCMC, bias increased as number of items increased beyond i = 5 (d = 0.87). Sample size was not found to impact the performance of ML, PLS, GSCA, or MCMC in the recovery of structural model parameter estimates.

Mean Differences of Standard Error Estimates for Measurement Models

Recovery of standard errors for the measurement model parameters was evaluated in terms of MAD between the standard error estimates and the empirical standard errors. In the overall MANOVA conducted for this study, the simple effect of estimation method on measurement model MAD was found to be large ($F(3) = 664127.05$, $p < .001$, partial $\eta^2 = 0.99$). Large interactions were also identified between estimation method and sample size ($F(6) = 696.66$, $p < .001$, partial $\eta^2 = 0.17$), between estimation method and number of items per latent variable ($F(6) = 1424279.72$, $p < .001$, partial $\eta^2 = 1.00$), and between estimation method and nature of the latent variable-indicator ($F(3) = 34366.03$, $p < .001$, partial $\eta^2 = 0.83$). Pair wise comparisons of the estimation methods revealed no differences in mean absolute differences for measurement model estimates between ML, PLS, and GSCA under conditions of correct specification and reflective latent variable-indicator relationships, correct specification and formative latent variable-indicator

Table 5. Mean average differences for measurement model standard errors by estimation method and experimental condition

		n = 50			*n* = 300			*n* = 1000		
		i = 3	*i* = 5	*i* = 7	*i* = 3	*i* = 5	*i* = 7	*i* = 3	*i* = 5	*i* = 7
ML	Correct Specification									
	Reflective	.020 (*SD* = .008)	.016 (*SD* = 0.005)	.015 (*SD* = 0.003)	.003 (*SD* = 0.001)	.003 (*SD* = 0.001)	.002 (*SD* = 0.001)	.001 (*SD* = 0.000)	.001 (*SD* = 0.000)	.001 (*SD* = 0.000)
	Formative	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)
	Misspecification									
	Reflective	.137 (*SD* = 0.008)	.028 (*SD* = 0.004)	.022 (*SD* = 0.003)	.016 (*SD* = 0.001)	.007 (*SD* = 0.001)	.005 (*SD* = 0.000)	.008 (*SD* = 0.000)	.004 (*SD* = 0.000)	.002 (*SD* = 0.000)
	Formative	.006 (*SD* = 0.002)	.002 (*SD* = 0.001)	.002 (*SD* = 0.001)	.002 (*SD* = 0.001)	.001 (*SD* = 0.000)	.000 (*SD* = 0.000)	.001 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)
PLS	Correct Specification									
	Reflective	.050 (*SD* = 0.018)	.034 (*SD* = 0.012)	.027 (*SD* = 0.010)	.016 (*SD* = 0.003)	.009 (*SD* = 0.001)	.006 (*SD* = 0.001)	.009 (*SD* = 0.001)	.005 (*SD* = 0.000)	.003 (*SD* = 0.000)
	Formative	.000 (*SD* = 0.000)	.000 (*SD* = 0.001)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)
	Misspecification									
	Reflective	.163 (*SD* = 0.009)	.043 (*SD* = 0.008)	.032 (*SD* = 0.007)	.028 (*SD* = 0.002)	.013 (*SD* = 0.001)	.009 (*SD* = 0.001)	.015 (*SD* = 0.001)	.007 (*SD* = 0.000)	.005 (*SD* = 0.000)
	Formative	.011 (*SD* = 0.001)	.008 (*SD* = 0.001)	.007 (*SD* = 0.001)	.004 (*SD* = 0.000)	.003 (*SD* = 0.000)	.003 (*SD* = 0.000)	.002 (*SD* = 0.000)	.002 (*SD* = 0.000)	.001 (*SD* = 0.000)
GSCA	Correct Specification									
	Reflective	.055 (*SD* = 0.008)	.032 (*SD* = 0.013)	.068 (*SD* = 0.061)	.021 (*SD* = 0.001)	.010 (*SD* = 0.001)	.010 (*SD* = 0.002)	.011 (*SD* = 0.000)	.006 (*SD* = 0.000)	.005 (*SD* = 0.001)
	Formative	.001 (*SD* = 0.000)	.001 (*SD* = 0.000)	.011 (*SD* = 0.002)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.022 (*SD* = 0.001)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.003 (*SD* = 0.000)
	Misspecification									
	Reflective	.168 (*SD* = 0.007)	.043 (*SD* = 0.012)	.069 (*SD* = 0.001)	.030 (*SD* = 0.001)	.014 (*SD* = 0.001)	.012 (*SD* = 0.001)	.016 (*SD* = 0.000)	.008 (*SD* = 0.000)	.006 (*SD* = 0.001)
	Formative	.012 (*SD* = 0.001)	.007 (*SD* = 0.001)	.015 (*SD* = 0.002)	.005 (*SD* = 0.000)	.003 (*SD* = 0.000)	.006 (*SD* = 0.000)	.003 (*SD* = 0.000)	.002 (*SD* = 0.000)	.003 (*SD* = 0.000)
MCMC	Correct Specification									
	Reflective	.021 (*SD* = 0.009)	.017 (*SD* = 0.004)	.606 (*SD* = 0.029)	.004 (*SD* = 0.001)	.003 (*SD* = 0.001)	.665 (*SD* = 0.013)	.002 (*SD* = 0.000)	.002 (*SD* = 0.000)	.680 (*SD* = 0.007)
	Formative	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.999 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.999 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.999 (*SD* = 0.000)
	Misspecification									
	Reflective	.136 (*SD* = 0.008)	.026 (*SD* = 0.004)	.615 (*SD* = 0.033)	.016 (*SD* = 0.001)	.007 (*SD* = 0.001)	.681 (*SD* = 0.012)	.009 (*SD* = 0.001)	.004 (*SD* = 0.000)	.696 (*SD* = 0.007)
	Formative	.011 (*SD* = 0.003)	.003 (*SD* = 0.001)	.978 (*SD* = 0.004)	.004 (*SD* = 0.000)	.001 (*SD* = 0.000)	.982 (*SD* = 0.001)	.002 (*SD* = 0.001)	.000 (*SD* = 0.000)	.979 (*SD* = 0.005)

relationships, misspecified models with reflective measurement models, or misspecified models with formative measurement models (all $d < 0.50$). For the

comparison of MAD values associated with measurement model standard error estimates, results from ML, PLS, and GSCA were combined. The amount of MAD observed for the measurement model estimates by estimation method and experimental condition is depicted in Figure 10 and reported in Table 5.

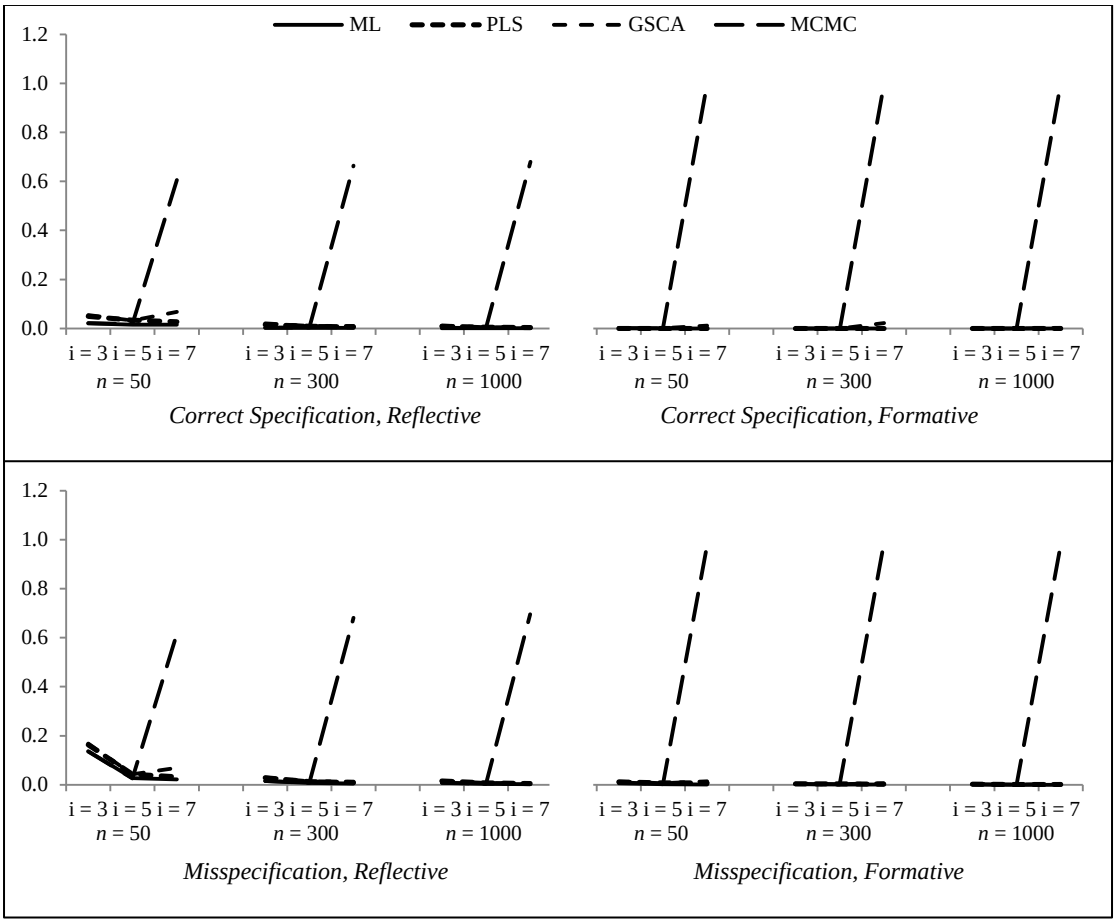


Figure 10. MAD of Measurement Model Standard Error Estimates.

Correct Specification, Reflective Indicators. Under conditions of correct model specification and reflective relationships between the latent variables and their indicators, large effects of number of items ($F(2) = 162659.74$, $p < .001$, partial $\eta^2 = 0.98$), and estimation method (ML, PLS, and GSCA *vs.* MCMC); $F(1) = 162055.81$, $p <$

.001, partial $\eta^2 = 0.97$) on measurement model MAD were observed. A large effect of sample size ($F(2) = 1350.30$, $p < .001$, partial $\eta^2 = 0.40$) was observed for the ML/PLS/GSCA group of methods, where MAD was observed to decrease as sample size increased. For MCMC, large effects of sample size ($F(2) = 174.12$, $p < .001$, partial $\eta^2 = 0.23$) and number of items ($F(2) = 560949.28$, $p < .001$, partial $\eta^2 = 1.00$) were found, as well as a large sample size × number of items interaction ($F(4) = 896.13$, $p < .001$, partial $\eta^2 = 0.75$). MAD of measurement model estimates yielded by MCMC were low across all levels of sample size and number of indicators per latent variable, except for conditions that include the largest number of items (i.e., seven indicators per latent variable), which consistently results in very high (i.e., close to 1.0) MAD values. Across all levels of sample size and number of indicators, ML, PLS, and GSCA recovered standard error estimates for the measurement model with MAD close to zero.

Correct Specification, Formative Indicators. Under conditions of correct model specification and formative relationships between the latent variables and their indicators, large effects of number of items ($F(2) = 8137356.94$, $p < .001$, partial $\eta^2 = 1.00$), and estimation method ($F(1) = 4709433.99$, $p < .001$, partial $\eta^2 = 1.00$) on measurement model MAD were observed, where estimation method consists of a comparison between ML/PLS/GSCA and MCMC. For MCMC, large effects were found for sample size ($F(2) = 500.16$, $p < .001$, partial $\eta^2 = 0.45$) and number of items per latent variable ($F(2) = 58685861505.00$, $p < .001$, partial $\eta^2 = 1.00$), and the sample size × number of items interaction ($F(4) = 4337.82$, $p < .001$, partial $\eta^2 = 0.94$). MAD of measurement model estimates yielded by MCMC were low across all levels of sample

size and number of indicators per latent variable, except for conditions that include the largest number of items (i.e., seven indicators per latent variable), which consistently results in very high (i.e., close to 1.0) MAD values.. Across all levels of sample size and number of indicators, ML, PLS, and GSCA recovered standard error estimates for the measurement model with MAD close to zero.

Misspecification, Reflective Indicators. Under conditions of misspecified models with reflective measurement models, large effects on MAD of measurement model standard error estimates were found for sample size ($F(2) = 8134.87$, $p < .001$, partial $\eta^2 = 0.75$), number of items ($F(2) = 377953.74$, $p < .001$, partial $\eta^2 = 0.99$), estimation method (i.e., ML/PLS.GSCA *vs.* MCMC; $F(1) = 410789.84$, $p < .001$, partial $\eta^2 = 0.99$), and the sample size × number of items × estimation method interaction ($F(4) = 992.08$, $p < .001$, partial $\eta^2 = 0.43$). Large effects for ML/PLS/GSCA measurement model MAD were found for sample size ($F(2) = 20757.57$, $p < .001$, partial $\eta^2 = 0.91$) and number of items ($F(2) = 9855.30$, $p < .001$, partial $\eta^2 = 0.83$). A large sample size × number of items interaction was also identified ($F(4) = 5717.04$, $p < .001$, partial $\eta^2 = 0.85$). This indicates that for the ML/PLS/GSCA group of estimation methods, measurement model MAD decreased as sample size and number of items increased. For MCMC estimation conditions, measurement model MAD was also found to be effected by sample size ($F(2) = 624.13$, $p < .001$, partial $\eta^2 = 0.50$) and number of items ($F(2) = 454794.67$, $p < .001$, partial $\eta^2 = 1.00$), along with a sample size × number of items interaction ($F(4) = 3620.81$, $p < .001$, partial $\eta^2 = 0.92$). MAD of measurement model estimates yielded by MCMC were low across all levels of sample size and number of

indicators per latent variable, except for conditions that include the largest number of items (i.e., seven indicators per latent variable), which consistently results in very high (i.e., close to 1.0) MAD values. Across all levels of sample size and number of indicators, ML, PLS, and GSCA recovered standard error estimates for the measurement model with MAD close to zero.

Misspecification, Formative Indicators. Under conditions of misspecified models with formative measurement models, moderate and large effects on MAD of measurement model standard error estimates were found for sample size ($F(2) = 228.53$, $p < .001$, partial $\eta^2 = 0.09$), number of items ($F(2) = 11659429.53$, $p < .001$, partial $\eta^2 = 1.00$), estimation method (i.e., ML/PLS.GSCA *vs.* MCMC; $F(1) = 2563989.01$, $p < .001$, partial $\eta^2 = 1.00$), and the sample size × number of items × estimation method interaction ($F(4) = 101.86$, $p < .001$, partial $\eta^2 = 0.08$). For both the ML/PLS/GSCA group of estimation methods and MCMC, moderate and large effects of sample size ($F(2) = 2666.91$, $p < .001$, partial $\eta^2 = 0.59$ and $F(2) = 55.28$, $p < .001$, partial $\eta^2 = 0.10$, respectively) and number of items ($F(2) = 399.33$, $p < .001$, partial $\eta^2 = 0.18$ and $F(2) = 15269106.62$, $p < .001$, partial $\eta^2 = 1.00$, respectively) were identified, as well as a moderate sample size × number of items interaction ($F(4) = 100.32$, $p < .001$, partial $\eta^2 = 0.10$ and $F(4) = 227.10$, $p < .001$, partial $\eta^2 = 0.47$, respectively for ML/PLS/GSCA and MCMC). Across estimation methods, measurement model MAD decreased as sample size and number of items decreased, except in the case of MCMC standard error estimates for models that include seven indicators per latent variable. MAD of measurement model estimates yielded by MCMC were low across all levels of sample

size and number of indicators per latent variable, except for conditions that include the largest number of items (i.e., seven indicators per latent variable), which consistently results in very high (i.e., close to 1.0) MAD values. Across all levels of sample size and number of indicators, ML, PLS, and GSCA recovered standard error estimates for the measurement model with MAD close to zero.

Mean Differences of Standard Error Estimates for Structural Models

Recovery of standard errors for the structural model parameters was evaluated in terms of the mean absolute difference (MAD) between the standard error estimates and

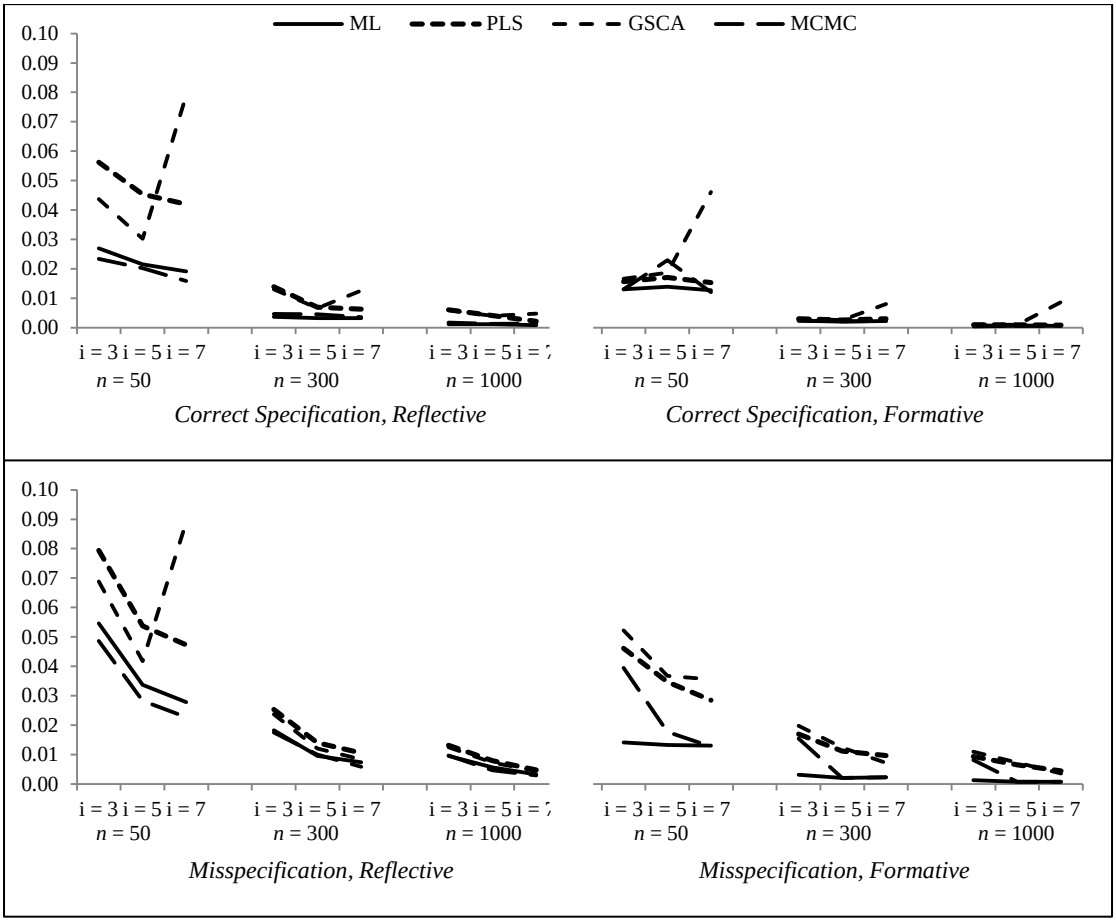


Figure 11. MAD of Structural Model Standard Error Estimates.

the empirical standard errors. In the overall MANOVA conducted for this study, the simple effect of estimation method on structural model MAD was found to be large (F(3)

Table 6. Mean average differences for structural model standard error estimates by estimation method and experimental condition

		n = 50			n = 300			n = 1000		
		i = 3	i = 5	i = 7	i = 3	i = 5	i = 7	i = 3	i = 5	i = 7
ML	Correct Specification									
	Reflective	.027 (*SD* = 0.013)	.022 (*SD* = 0.013)	.019 (*SD* = 0.010)	.004 (*SD* = 0.002)	.003 (*SD* = 0.002)	.003 (*SD* = 0.002)	.001 (*SD* = 0.001)	.001 (*SD* = 0.001)	.001 (*SD* = 0.000)
	Formative	.013 (*SD* = 0.007)	.014 (*SD* = 0.008)	.013 (*SD* = 0.007)	.002 (*SD* = 0.001)	.002 (*SD* = 0.001)	.002 (*SD* = 0.001)	.000 (*SD* = 0.000)	.001 (*SD* = 0.000)	.001 (*SD* = 0.000)
	Misspecification									
	Reflective	.055 (*SD* = 0.017)	.034 (*SD* = 0.014)	.028 (*SD* = 0.012)	.018 (*SD* = 0.003)	.010 (*SD* = 0.002)	.007 (*SD* = 0.003)	.010 (*SD* = 0.001)	.006 (*SD* = 0.001)	.003 (*SD* = 0.001)
	Formative	.014 (*SD* = 0.008)	.013 (*SD* = 0.007)	.013 (*SD* = 0.007)	.003 (*SD* = 0.002)	.002 (*SD* = 0.001)	.002 (*SD* = 0.001)	.001 (*SD* = 0.001)	.001 (*SD* = 0.000)	.001 (*SD* = 0.000)
PLS	Correct Specification									
	Reflective	.056 (*SD* = 0.017)	.045 (*SD* = 0.015)	.042 (*SD* = 0.014)	.014 (*SD* = 0.003)	.007 (*SD* = 0.002)	.006 (*SD* = 0.003)	.006 (*SD* = 0.001)	.004 (*SD* = 0.001)	.002 (*SD* = 0.001)
	Formative	.016 (*SD* = 0.008)	.017 (*SD* = 0.010)	.015 (*SD* = 0.008)	.003 (*SD* = 0.002)	.003 (*SD* = 0.002)	.003 (*SD* = 0.002)	.001 (*SD* = 0.000)	.001 (*SD* = 0.000)	.001 (*SD* = 0.000)
	Misspecification									
	Reflective	.079 (*SD* = 0.014)	.054 (*SD* = 0.012)	.047 (*SD* = 0.011)	.025 (*SD* = 0.003)	.014 (*SD* = 0.002)	.011 (*SD* = 0.003)	.013 (*SD* = 0.001)	.008 (*SD* = 0.001)	.005 (*SD* = 0.001)
	Formative	.046 (*SD* = 0.008)	.035 (*SD* = 0.009)	.028 (*SD* = 0.008)	.017 (*SD* = 0.002)	.011 (*SD* = 0.002)	.010 (*SD* = 0.002)	.009 (*SD* = 0.000)	.007 (*SD* = 0.001)	.004 (*SD* = 0.001)
GSCA	Correct Specification									
	Reflective	.044 (*SD* = 0.016)	.030 (*SD* = 0.016)	.079 (*SD* = 0.087)	.013 (*SD* = 0.003)	.007 (*SD* = 0.003)	.013 (*SD* = 0.006)	.006 (*SD* = 0.001)	.004 (*SD* = 0.001)	.005 (*SD* = 0.002)
	Formative	.017 (*SD* = 0.009)	.019 (*SD* = 0.011)	.046 (*SD* = 0.025)	.003 (*SD* = 0.002)	.003 (*SD* = 0.002)	.008 (*SD* = 0.002)	.001 (*SD* = 0.001)	.001 (*SD* = 0.001)	.009 (*SD* = 0.002)
	Misspecification									
	Reflective	.069 (*SD* = 0.017)	.042 (*SD* = 0.020)	.089 (*SD* = 0.002)	.024 (*SD* = 0.003)	.012 (*SD* = 0.003)	.008 (*SD* = 0.005)	.013 (*SD* = 0.001)	.007 (*SD* = 0.001)	.004 (*SD* = 0.002)
	Formative	.052 (*SD* = 0.010)	.037 (*SD* = 0.011)	.036 (*SD* = 0.021)	.020 (*SD* = 0.002)	.012 (*SD* = 0.002)	.007 (*SD* = 0.003)	.011 (*SD* = 0.001)	.007 (*SD* = 0.001)	.004 (*SD* = 0.001)
MCMC	Correct Specification									
	Reflective	.023 (*SD* = 0.013)	.020 (*SD* = 0.011)	.016 (*SD* = 0.009)	.005 (*SD* = 0.002)	.005 (*SD* = 0.002)	.004 (*SD* = 0.002)	.002 (*SD* = 0.001)	.001 (*SD* = 0.001)	.001 (*SD* = 0.001)
	Formative	.013 (*SD* = 0.007)	.023 (*SD* = 0.010)	.012 (*SD* = 0.006)	.002 (*SD* = 0.001)	.002 (*SD* = 0.001)	.002 (*SD* = 0.001)	.001 (*SD* = 0.000)	.001 (*SD* = 0.000)	.001 (*SD* = 0.000)
	Misspecification									
	Reflective	.049 (*SD* = 0.017)	.028 (*SD* = 0.013)	.023 (*SD* = 0.011)	.018 (*SD* = 0.003)	.010 (*SD* = 0.003)	.006 (*SD* = 0.003)	.010 (*SD* = 0.001)	.005 (*SD* = 0.001)	.003 (*SD* = 0.001)
	Formative	.039 (*SD* = 0.015)	.018 (*SD* = 0.016)	.013 (*SD* = 0.006)	.015 (*SD* = 0.002)	.002 (*SD* = 0.001)	.002 (*SD* = 0.001)	.008 (*SD* = 0.006)	.001 (*SD* = 0.000)	.001 (*SD* = 0.000)

= 1226.71, $p < .001$, partial $\eta^2 = 0.15$); and moderate effects were found for model misspecification ($F(1) = 1670.08$, $p < .001$, partial $\eta^2 = 0.08$) and type of latent variable-indicator relationships ($F(1) = 2174.70$, $p < .001$, partial $\eta^2 = 0.10$). A moderate interaction was also identified between estimation method and sample size ($F(6) = 368.59$, $p < .001$, partial $\eta^2 = 0.10$), as well as between sample size, number of items, and estimation method ($F(12) = 150.35$, $p < .001$, partial $\eta^2 = 0.08$). The amount of MAD observed for the structural model estimates by estimation method and experimental condition is depicted in Figure 11 and presented in Table 6.

Correct Specification, Reflective Indicators. Under conditions of correct model specification with reflective indicators, large and moderate effects were found for sample size ($F(2) = 1872.56$, $p < .001$, partial $\eta^2 = 0.42$) and estimation method ($F(3) = 248.69$, $p < .001$, partial $\eta^2 = 0.13$). Pair wise comparisons across estimation methods indicated no difference in the amount of MAD observed for the structural model parameter estimates between PLS and GSCA estimation methods ($p > .05$). For the remainder of analyses under this condition for this outcome, the results from PLS and GSCA were combined. A large effect of sample size on MAD for structural model estimates was observed for ML ($F(2) = 1268.08$, $p < .001$, partial $\eta^2 = 0.66$), PLS/GSCA ($F(2) = 1006.09$, $p < .001$, partial $\eta^2 = 0.43$), and MCMC ($F(2) = 858.21$, $p < .001$, partial $\eta^2 = 0.59$). Interestingly, GSCA resulted in smaller MAD for $i = 5$ than $i = 3$ or 7 when $n = 50$. However, the difference in MAD values across levels of i for the smallest sample size is quite small. For all estimation methods, structural model MAD was found to

decrease as sample size increased. Averaged across all models, MAD was found to be highest for PLS and GSCA, and lowest for ML and MCMC.

Correct Specification, Formative Indicators. Under conditions of correct model specification with formative indicators, large effects were found for sample size ($F(2) = 2307.83$, $p < .001$, partial $\eta^2 = 0.47$) and estimation method ($F(2) = 273.99$, $p < .001$, partial $\eta^2 = 0.14$). A moderate sample size × estimation method interaction was also found ($F(6) = 92.30$, $p < .001$, partial $\eta^2 = 0.10$). For both ML and PLS, a large effect of sample size ($F(2) = 1177.62$, $p < .001$, partial $\eta^2 = 0.64$ and $F(2) = 1109.09$, $p < .001$, partial $\eta^2 = 0.62$, respectively) indicated a decrease in MAD for the structural model standard error estimates as sample size increased. For both GSCA and MCMC, large effects were found for sample size (GSCA: $F(2) = 838.65$, $p < .001$, partial $\eta^2 = 0.56$; MCMC: $F(2) = 748.89$, $p < .001$, partial $\eta^2 = 0.56$), number of items (GSCA: $F(2) = 299.04$, $p < .001$, partial $\eta^2 = 0.31$; MCMC: $F(2) = 40.32$, $p < .001$, partial $\eta^2 = 0.06$), and the sample size × number of items interaction (GSCA: $F(4) = 85.56$, $p < .001$, partial $\eta^2 = 0.20$; MCMC: $F(4) = 25.43$, $p < .001$, partial $\eta^2 = 0.08$). For GSCA, MAD for structural model estimates decreased as sample size increased, but increased as the number of items increased from $i = 5$ to $i = 7$ within each level of sample size. For MCMC, MAD for structural model estimates decreased as sample size and number of indicators per latent variable increased, except in the case of the smallest ($n = 50$) sample size, for which MAD was greater for models that included seven items than for any other combination of sample size and number of indicators per latent variable.

Misspecification, Reflective Indicators. Under conditions of misspecified models with reflective indicators, large effects on the MAD values associated with structural model estimates were found for sample size ($F(2) = 13416.00$, $p < .001$, partial $\eta^2 = 0.84$), number of items ($F(2) = 1365.79$, $p < .001$, partial $\eta^2 = 0.34$), and estimation method ($F(3) = 855.40$, $p < .001$, partial $\eta^2 = 0.33$). Additionally, the sample size × number of items × estimation method interaction effect was found to be large ($F(12) = 165.15$, $p < .001$, partial $\eta^2 = 0.27$). Pair wise comparisons across estimation methods indicated no difference in the amount of MAD observed for the structural model parameter estimates between PLS and GSCA estimation methods within the context of misspecified models with reflective measurement models. For ML estimation, large effects were observed for sample size ($F(2) = 1846.84$, $p < .001$, partial $\eta^2 = 0.74$) and number of items ($F(2) = 358.52$, $p < .001$, partial $\eta^2 = 0.35$), as well as a significant sample size × number of items interaction ($F(4) = 65.52$, $p < .001$, partial $\eta^2 = 0.17$). For the ML approach, MAD associated with the structural model was found to decrease as sample size and number of items increase.

For both PLS/GSCA and MCMC, large effects were observed for sample size (PLS/GSCA: $F(2) = 6531.28$, $p < .001$, partial $\eta^2 = 0.83$; MCMC: $F(2) = 1312.52$, $p < .001$, partial $\eta^2 = 0.67$), number of items (PLS/GSCA: $F(2) = 394.60$, $p < .001$, partial $\eta^2 = 0.23$; MCMC: $F(2) = 395.75$, $p < .001$, partial $\eta^2 = 0.39$), and the sample size × number of items interaction (PLS/GSCA: $F(4) = 133.29$, $p < .001$, partial $\eta^2 = 0.17$; MCMC: $F(4) = 62.16$, $p < .001$, partial $\eta^2 = 0.16$). For PLS/GSCA and MCMC estimation methods, MAD associated with structural model estimates decreased as sample size and number of

items increased, with the least amount of change across levels of *i* occurring when the sample was at its largest (i.e., *n* = 1000). Overall, mean MAD values were largest for PLS/GSCA parameter estimates and smallest for MCMC and ML parameter estimates.

Misspecification, Formative Indicators. Under conditions of misspecified models with formative measurement models, large effects on MAD associated with structural model parameter estimates were found for sample size ($F(2) = 1468.00$, $p < .001$, partial $\eta^2 = 0.38$), number of items ($F(2) = 754.81$, $p < .001$, partial $\eta^2 = 0.24$), and estimation method ($F(3) = 971.16$, $p < .001$, partial $\eta^2 = 0.38$). A large effect of sample size ($F(2) = 601.13$, $p < .001$, partial $\eta^2 = 0.53$) was found for MAD of structural model estimates recovered by ML, where MAD decreased as sample size increased. For PLS, large effects were found for sample size ($F(2) = 4382.57$, $p < .001$, partial $\eta^2 = 0.87$), number of items ($F(2) = 456.90$, $p < .001$, partial $\eta^2 = 0.41$), and the sample size × number of items interaction ($F(4) = 71.18$, $p < .001$, partial $\eta^2 = 0.18$). For GSCA, large effects were found for sample size ($F(2) = 88.48$, $p < .001$, partial $\eta^2 = 0.15$), number of items ($F(2) = 510.08$, $p < .001$, partial $\eta^2 = 0.50$), and the sample size × number of items interaction ($F(4) = 75.47$, $p < .001$, partial $\eta^2 = 0.23$). GSCA resulted in smaller MAD for $i = 5$ than $i = 3$ or 7 when $n = 50$, but the differences between these MAD values were quite small. For MCMC, large effects were found for sample size ($F(2) = 1998.83$, $p < .001$, partial $\eta^2 = 0.75$), number of items ($F(2) = 238.60$, $p < .001$, partial $\eta^2 = 0.26$), and the sample size × number of items interaction ($F(4) = 20.44$, $p < .001$, partial $\eta^2 = 0.06$). For PLS, GSCA, and MCMC, MAD decreased as sample size and number of items increased, with the greatest change observed as number of items increased within $n = 50$.

Accuracy of Standard Error Estimates for Measurement Models

The accuracy with which standard errors were recovered for the measurement model was evaluated by constructing a confidence interval of ± 1.96 standard errors around each parameter estimate and determining whether the corresponding population parameter fell within its bounds. The outcome variable of interest is the proportion of measurement model parameter estimates for which the standard errors were found to be accurate. In the overall MANOVA conducted for this study, the simple effect of estimation method on measurement model accuracy was found to be large ($F(3)$ =

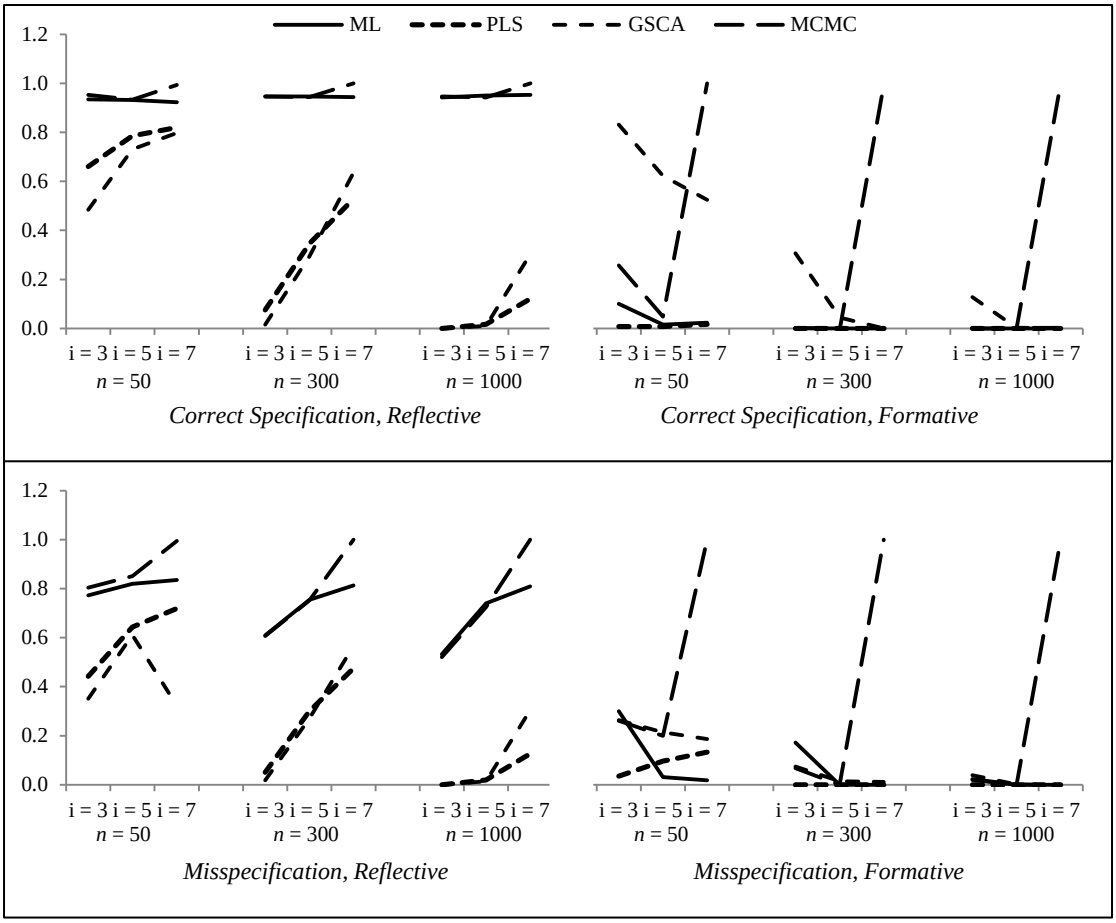


Figure 12. Accuracy of Measurement Model Estimates.

Table 7. Mean accuracy of measurement model estimates by estimation method and experimental condition

		n = 50			*n* = 300			*n* = 1000		
		i = 3	*i* = 5	*i* = 7	*i* = 3	*i* = 5	*i* = 7	*i* = 3	*i* = 5	*i* = 7
ML	Correct Specification									
	Reflective	.934 (*SD* = 0.085)	.932 (*SD* = 0.072)	.923 (*SD* = 0.057)	.947 (*SD* = 0.080)	.946 (*SD* = 0.058)	.943 (*SD* = 0.058)	.943 (*SD* = 0.076)	.950 (*SD* = 0.063)	.953 (*SD* = 0.051)
	Formative	.100 (*SD* = 0.093)	.016 (*SD* = 0.036)	.024 (*SD* = 0.039)	.001 (*SD* = 0.009)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)
	Misspecification									
	Reflective	.773 (*SD* = 0.116)	.820 (*SD* = 0.088)	.836 (*SD* = 0.071)	.609 (*SD* = 0.090)	.755 (*SD* = 0.056)	.813 (*SD* = 0.047)	.532 (*SD* = 0.106)	.740 (*SD* = 0.070)	.810 (*SD* = 0.052)
	Formative	.300 (*SD* = 0.071)	.031 (*SD* = 0.046)	.018 (*SD* = 0.033)	.172 (*SD* = 0.056)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.020 (*SD* = 0.043)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)
PLS	Correct Specification									
	Reflective	.660 (*SD* = 0.209)	.785 (*SD* = 0.144)	.819 (*SD* = 0.114)	.076 (*SD* = 0.102)	.346 (*SD* = 0.155)	.529 (*SD* = 0.168)	.000 (*SD* = 0.000)	.017 (*SD* = 0.036)	.121 (*SD* = 0.097)
	Formative	.008 (*SD* = 0.047)	.007 (*SD* = 0.047)	.017 (*SD* = 0.071)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)
	Misspecification									
	Reflective	.442 (*SD* = 0.164)	.644 (*SD* = 0.124)	.719 (*SD* = 0.121)	.051 (*SD* = 0.075)	.300 (*SD* = 0.130)	.475 (*SD* = 0.150)	.000 (*SD* = 0.000)	.020 (*SD* = 0.042)	.127 (*SD* = 0.094)
	Formative	.035 (*SD* = 0.081)	.096 (*SD* = 0.120)	.132 (*SD* = 0.139)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)
GSCA	Correct Specification									
	Reflective	.484 (*SD* = 0.179)	.730 (*SD* = 0.140)	.798 (*SD* = 0.097)	.015 (*SD* = 0.042)	.294 (*SD* = 0.150)	.636 (*SD* = 0.151)	.000 (*SD* = 0.000)	.010 (*SD* = 0.028)	.304 (*SD* = 0.138)
	Formative	.831 (*SD* = 0.128)	.623 (*SD* = 0.142)	.524 (*SD* = 0.166)	.306 (*SD* = 0.129)	.044 (*SD* = 0.052)	.000 (*SD* = 0.000)	.128 (*SD* = 0.089)	.002 (*SD* = 0.012)	.002 (*SD* = 0.010)
	Misspecification									
	Reflective	.350 (*SD* = 0.159)	.611 (*SD* = 0.130)	.322 (*SD* = 0.144)	.019 (*SD* = 0.045)	.270 (*SD* = 0.127)	.567 (*SD* = 0.122)	.000 (*SD* = 0.000)	.015 (*SD* = 0.034)	.304 (*SD* = 0.138)
	Formative	.264 (*SD* = 0.082)	.214 (*SD* = 0.077)	.186 (*SD* = 0.090)	.075 (*SD* = 0.055)	.014 (*SD* = 0.027)	.012 (*SD* = 0.021)	.039 (*SD* = 0.053)	.000 (*SD* = 0.005)	.000 (*SD* = 0.000)
MCMC	Correct Specification									
	Reflective	.953 (*SD* = 0.073)	.933 (*SD* = 0.065)	.993 (*SD* = 0.017)	.946 (*SD* = 0.075)	.944 (*SD* = 0.058)	1.000 (*SD* = 0.000)	.946 (*SD* = 0.069)	.943 (*SD* = 0.065)	1.000 (*SD* = 0.000)
	Formative	.256 (*SD* = 0.144)	.048 (*SD* = 0.052)	1.000 (*SD* = 0.000)	.001 (*SD* = 0.009)	.000 (*SD* = 0.000)	1.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	1.000 (*SD* = 0.000)
	Misspecification									
	Reflective	.804 (*SD* = 0.113)	.851 (*SD* = 0.090)	.995 (*SD* = 0.015)	.607 (*SD* = 0.091)	.751 (*SD* = 0.059)	1.000 (*SD* = 0.000)	.521 (*SD* = 0.102)	.724 (*SD* = 0.078)	1.000 (*SD* = 0.000)
	Formative	.260 (*SD* = .129)	.200 (*SD* = 0.094)	1.000 (*SD* = 0.000)	.068 (*SD* = 0.091)	.000 (*SD* = 0.000)	1.000 (*SD* = 0.000)	.023 (*SD* = 0.051)	.000 (*SD* = 0.000)	1.000 (*SD* = 0.000)

17379.03, $p < .001$, partial $\eta^2 = 0.72$). Moderate and large interactions were identified between estimation method and sample size ($F(6) = 1674.34$, $p < .001$, partial $\eta^2 = 0.33$),

number of items per latent variable ($F(6) = 3767.15$, $p < .001$, partial $\eta^2 = 0.53$), degree of model misspecification ($F(3) = 475.19$, $p < .001$, partial $\eta^2 = 0.07$), and nature of the latent variable-indicator ($F(3) = 14326.05$, $p < .001$, partial $\eta^2 = 0.68$). The degree of accuracy of the measurement model estimates are presented by estimation method and experimental condition in Figure 12 and Table 7.

Correct Specification, Reflective Indicators. Pair-wise comparisons yielded no differences in accuracy of measurement model estimates between ML and MCMC or between PLS and GSCA under conditions of correctly specified models with reflective indicators. No effects were found for measurement model accuracy of ML/MCMC estimates under conditions of correct model specification and reflective indicators. Across all levels of sample size and number of indicators, ML/MCMC produced estimates with consistently high levels of accuracy (M = 0.9448, 0.9416, and .09663, for 3, 5, and 7 items, respectively). For PLS and GSCA methods, large effects were found for sample size ($F(2) = 5223.75$, $p < .001$, partial $\eta^2 = 0.80$) and number of items ($F(2) = 1364.28$, $p < .001$, partial $\eta^2 = 0.50$), as well as a large sample size × number of items interaction ($F(4) = 174.09$, $p < .001$, partial $\eta^2 = 0.21$). For PLS and GSCA, the accuracy of measurement model estimates increased as number of items increased within a sample size, but decreased as sample size increased.

Correct Specification, Formative Indicators. Pair-wise comparisons yielded no differences in accuracy of measurement model estimates between ML and PLS (p = 1.00) under conditions of correct model specification and formative latent variable-indicator relationships. Under conditions of correctly specified models with formative

measurement models, large effects on accuracy of measurement model estimates were found for sample size ($F(2) = 3652.15$, $p < .001$, partial $\eta^2 = 0.58$), number of items ($F(2) = 7559.96$, $p < .001$, partial $\eta^2 = 0.74$), and estimation method (i.e., ML/PLS *vs.* GSCA *vs.* MCMC; $F(2) = 12436.02$, $p < .001$, partial $\eta^2 = 0.83$). A moderate interaction between estimation method, sample size, and number of items was also identified ($F(8) = 89.12$, $p < .001$, partial $\eta^2 = 0.12$). A moderate effect of sample size on accuracy of measurement model estimates was found for ML/PLS ($F(2) = 168.20$, $p < .001$, partial $\eta^2 = 0.11$), where accuracy of parameter estimates was found to decrease as sample size increased. Across all levels of sample size and number of indicators per latent variable, both ML and PLS produced very few accurate estimates for the measurement model under conditions of formative measurement models paired with correctly specified models.

Large effects of sample size ($F(2) = 5007.44$, $p < .001$, partial $\eta^2 = 0.88$) and number of items per latent variable ($F(2) = 755.44$, $p < .001$, partial $\eta^2 = 0.53$) were found for GSCA estimates, in addition to a moderate sample size × number of items interaction ($F(4) = 44.26$, $p < .001$, partial $\eta^2 = 0.12$). The accuracy of GSCA parameter estimates for the measurement model was found to decrease and sample size and number of items increased. For MCMC estimation, large effects of sample size ($F(2) = 154.60$, $p < .001$, partial $\eta^2 = 0.21$) and number of items ($F(2) = 39980.97$, $p < .001$, partial $\eta^2 = 0.99$) were observed, as well as a large sample size × number of items interaction effect ($F(4) = 317.62$, $p < .001$, partial $\eta^2 = 0.51$). The accuracy of MCMC measurement model parameter estimates was found to decrease as sample size and number of items increased, except in instances where $i = 7$. The accuracy of MCMC measurement model parameter

estimates was also found to be lower when $i = 5$ than when $i = 3$ or 7 when $n = 50$. Under conditions that included seven items per latent variable, 100% of MCMC estimates were accurate.

Misspecification, Reflective Indicators. Large effects of sample size ($F(2) = 3677.00$, $p < .001$, partial $\eta^2 = 0.58$), number of items per latent variable ($F(2) = 3457.14$, $p < .001$, partial $\eta^2 = 0.57$), and estimation method ($F(3) = 11010.90$, $p < .001$, partial $\eta^2 = 0.86$) were found for accuracy of measurement model parameters recovered within the context of misspecified reflective models. Despite pair wise comparisons indicating differences between the four estimation methods, the same pattern of results emerged across all four methods. Specifically, the effects of sample size (ML: $F(2) = 244.38$, $p < .001$, partial $\eta^2 = 0.27$; PLS: $F(2) = 2782.43$, $p < .001$, partial $\eta^2 = 0.81$; GSCA: $F(2) = 904.39$, $p < .001$, partial $\eta^2 = 0.57$; MCMC: $F(2) = 340.34$, $p < .001$, partial $\eta^2 = 0.35$) and number of items per latent variable (ML: $F(2) = 608.58$, $p < .001$, partial $\eta^2 = 0.48$; PLS: $F(2) = 689.37$, $p < .001$, partial $\eta^2 = 0.51$; GSCA: $F(2) = 674.32$, $p < .001$, partial $\eta^2 = 0.50$; MCMC: $F(2) = 2353.01$, $p < .001$, partial $\eta^2 = 0.79$) were found to be large, as well as the sample × number it items interaction effect (ML: $F(4) = 73.30$, $p < .001$, partial $\eta^2 = 0.18$; PLS: $F(4) = 78.86$, $p < .001$, partial $\eta^2 = 0.19$; GSCA: $F(4) = 419.67$, $p < .001$, partial $\eta^2 = 0.56$; MCMC: $F(4) = 130.15$, $p < .001$, partial $\eta^2 = 0.29$).

Across all estimation methods, accuracy of the measurement model estimates was found to increase within each level of sample size as the number of items increased. However, accuracy of estimated was found to decrease as sample size increased, except in the case of MCMC estimates for $i = 7$, which were found to be 100% accurate.

Although the pattern of results was found to be consistent across methods, ML and MCMC consistently yielded more accurate estimates than either PLS or GSCA, and PLS yielded more accurate estimates than GSCA. It is worth noting that accuracy of measurement model estimates for GSCA with $n = 50$ was higher for 5 items than for 3 or 7 items.

Misspecification, Formative Indicators. Pair wise comparisons yielded no differences in accuracy of measurement model estimates between ML and PLS ($p = 1.00$) under conditions of model misspecification and formative latent variable-indicator relationships. Moderate and large effects were found for sample size ($F(2) = 337.88$, $p < .001$, partial $\eta^2 = 0.12$), number of items ($F(2) = 6960.35$, $p < .001$, partial $\eta^2 = 0.74$), and estimation method ($F(2) = 2444.40$, $p < .001$, partial $\eta^2 = 0.51$). A large effect of sample size ($F(2) = 218.26$, $p < .001$, partial $\eta^2 = 0.15$) was observed for the ML/PLS approaches, where accuracy of the measurement model estimates were found to decrease as sample size increased. For GSCA, large effects were found for both sample size ($F(2) = 1907.47$, $p < .001$, partial $\eta^2 = 0.74$) and number of items ($F(2) = 148.77$, $p < .001$, partial $\eta^2 = 0.18$), which indicated that measurement model accuracy decreased as sample size increased as well as when number of items increased.

Similarly to GSCA, moderate and large effects of sample size ($F(2) = 48.60$, $p < .001$, partial $\eta^2 = 0.09$) and number of items ($F(2) = 18664.88$, $p < .001$, partial $\eta^2 = 0.97$) were found for the MCMC approach, where measurement model accuracy decreased as sample size and number of items increased. However, a large sample size × number of items interaction ($F(4) = 139.87$, $p < .001$, partial $\eta^2 = 0.35$) was also found for MCMC.

Although measurement model accuracy for MCMC under conditions of model misspecification and formative indicators decreased as sample size and number of items increased, MCMC produced estimates with 100% accuracy when $i = 7$ for all sample sizes.

Accuracy of Standard Error Estimates for Structural Models

The accuracy with which standard errors were recovered for the structural model was evaluated by constructing a confidence interval of ± 1.96 standard errors around each parameter estimate and determining whether the corresponding population parameter fell

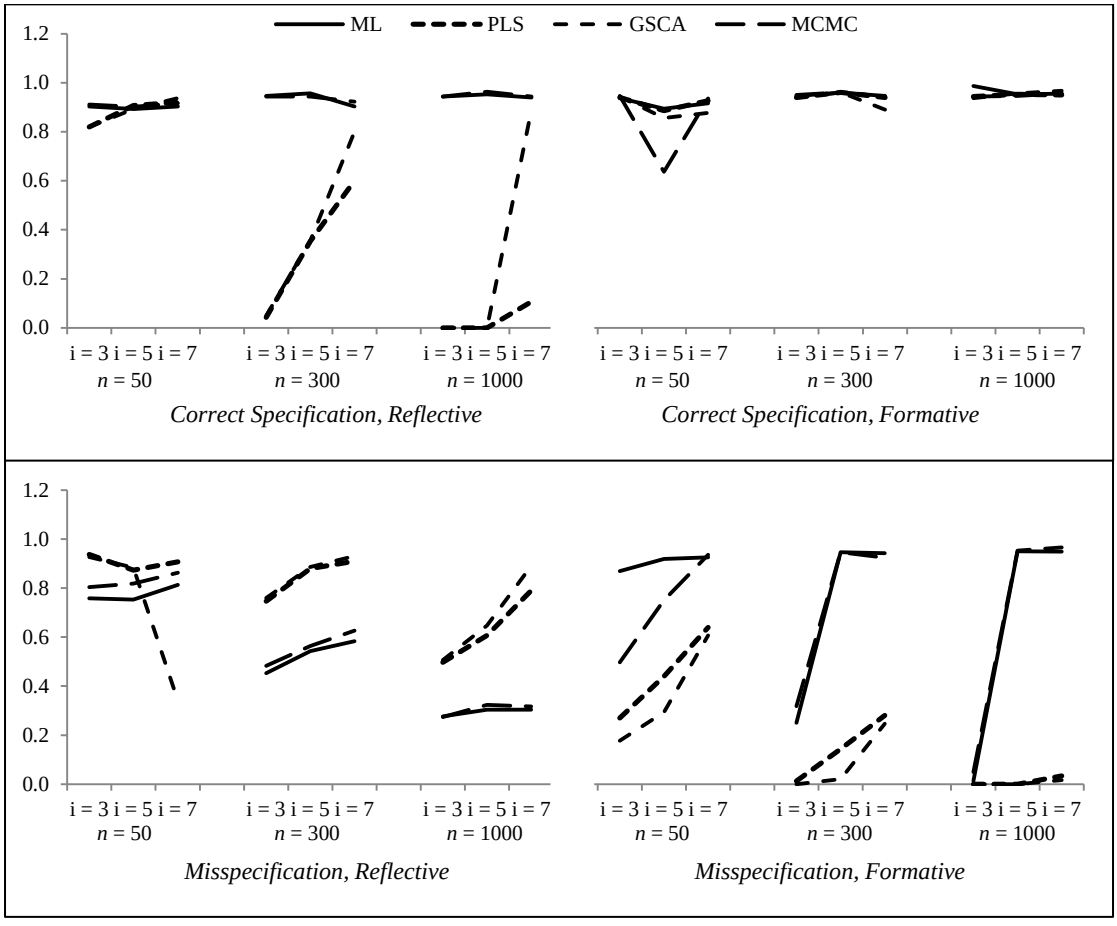


Figure 13. Accuracy of Structural Model Estimates.

Table 8. Mean accuracy of structural model estimates by estimation method and experimental condition

		n = 50			*n* = 300			*n* = 1000		
		i = 3	*i* = 5	*i* = 7	*i* = 3	*i* = 5	*i* = 7	*i* = 3	*i* = 5	*i* = 7
ML	Correct Specification									
	Reflective	.903 (*SD* = 0.198)	.893 (*SD* = 0.214)	.903 (*SD* = 0.198)	.947 (*SD* = 0.165)	.957 (*SD* = 0.141)	.903 (*SD* = 0.214)	.943 (*SD* = 0.169)	.953 (*SD* = 0.146)	.940 (*SD* = 0.163)
	Formative	.937 (*SD* = 0.177)	.894 (*SD* = 0.236)	.917 (*SD* = 0.197)	.947 (*SD* = 0.155)	.959 (*SD* = 0.149)	.946 (*SD* = 0.156)	.987 (*SD* = 0.081)	.953 (*SD* = 0.146)	.955 (*SD* = 0.143)
	Misspecification									
	Reflective	.759 (*SD* = 0.278)	.753 (*SD* = 0.311)	.813 (*SD* = 0.275)	.453 (*SD* = 0.241)	.543 (*SD* = 0.252)	.583 (*SD* = 0.345)	.277 (*SD* = 0.249)	.303 (*SD* = 0.245)	.303 (*SD* = 0.265)
	Formative	.870 (*SD* = 0.270)	.918 (*SD* = 0.203)	.925 (*SD* = 0.199)	.250 (*SD* = 0.341)	.946 (*SD* = 0.155)	.942 (*SD* = 0.171)	.013 (*SD* = 0.079)	.950 (*SD* = 0.151)	.949 (*SD* = 0.152)
PLS	Correct Specification									
	Reflective	.820 (*SD* = 0.285)	.907 (*SD* = 0.204)	.920 (*SD* = 0.193)	.043 (*SD* = 0.141)	.353 (*SD* = 0.340)	.600 (*SD* = 0.375)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.107 (*SD* = 0.221)
	Formative	.940 (*SD* = 0.182)	.887 (*SD* = 0.247)	.927 (*SD* = 0.187)	.940 (*SD* = 0.173)	.960 (*SD* = 0.148)	.940 (*SD* = 0.163)	.943 (*SD* = 0.169)	.953 (*SD* = 0.146)	.950 (*SD* = 0.151)
	Misspecification									
	Reflective	.937 (*SD* = 0.167)	.873 (*SD* = 0.254)	.907 (*SD* = 0.204)	.747 (*SD* = 0.276)	.880 (*SD* = 0.222)	.910 (*SD* = 0.193)	.497 (*SD* = 0.159)	.607 (*SD* = 0.269)	.790 (*SD* = 0.261)
	Formative	.270 (*SD* = 0.250)	.440 (*SD* = 0.283)	.640 (*SD* = 0.358)	.013 (*SD* = 0.081)	.143 (*SD* = 0.227)	.280 (*SD* = 0.256)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.033 (*SD* = 0.125)
GSCA	Correct Specification									
	Reflective	.820 (*SD* = 0.297)	.893 (*SD* = 0.221)	.937 (*SD* = 0.186)	.050 (*SD* = 0.151)	.357 (*SD* = 0.354)	.800 (*SD* = 0.307)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.890 (*SD* = 0.216)
	Formative	.943 (*SD* = 0.179)	.857 (*SD* = 0.268)	.877 (*SD* = 0.231)	.940 (*SD* = 0.173)	.960 (*SD* = 0.148)	.890 (*SD* = 0.216)	.937 (*SD* = 0.186)	.957 (*SD* = 0.141)	.967 (*SD* = 0.125)
	Misspecification									
	Reflective	.927 (*SD* = 0.177)	.883 (*SD* = 0.235)	.340 (*SD* = 0.334)	.760 (*SD* = 0.270)	.887 (*SD* = 0.225)	.930 (*SD* = 0.174)	.507 (*SD* = 0.164)	.647 (*SD* = 0.263)	.890 (*SD* = 0.216)
	Formative	.177 (*SD* = 0.240)	.293 (*SD* = 0.279)	.607 (*SD* = 0.331)	.000 (*SD* = 0.000)	.020 (*SD* = 0.098)	.247 (*SD* = 0.257)	.000 (*SD* = 0.000)	.000 (*SD* = 0.000)	.017 (*SD* = 0.090)
MCMC	Correct Specification									
	Reflective	.912 (*SD* = 0.200)	.902 (*SD* = 0.212)	.917 (*SD* = 0.188)	.943 (*SD* = 0.159)	.943 (*SD* = 0.159)	.923 (*SD* = 0.190)	.943 (*SD* = 0.179)	.963 (*SD* = 0.143)	.943 (*SD* = 0.159)
	Formative	.947 (*SD* = 0.165)	.636 (*SD* = 0.393)	.936 (*SD* = 0.177)	.950 (*SD* = 0.151)	.960 (*SD* = 0.148)	.943 (*SD* = 0.159)	.943 (*SD* = 0.169)	.947 (*SD* = 0.155)	.957 (*SD* = 0.141)
	Misspecification									
	Reflective	.803 (*SD* = 0.265)	.819 (*SD* = 0.288)	.863 (*SD* = 0.224)	.483 (*SD* = 0.227)	.563 (*SD* = 0.248)	.627 (*SD* = 0.334)	.273 (*SD* = 0.250)	.323 (*SD* = 0.247)	.317 (*SD* = 0.262)
	Formative	.497 (*SD* = 0.188)	.750 (*SD* = 0.354)	.935 (*SD* = 0.178)	.318 (*SD* = 0.257)	.946 (*SD* = 0.155)	.925 (*SD* = 0.194)	.050 (*SD* = 0.150)	.953 (*SD* = 0.146)	.966 (*SD* = 0.127)

within its bounds. The outcome variable of interest is the proportion of structural model parameter estimates for which the standard errors were found to be accurate. In the

overall MANOVA conducted for this study, the simple effect of estimation method on structural model accuracy was found to be large ($F(3) = 1153.60$, $p < .001$, partial $\eta^2 = 0.15$). A large interaction effect was also identified between estimation method, degree of misspecification, and the nature of the latent variable-indicator ($F(3) = 2888.76$, $p < .001$, partial $\eta^2 = 0.30$). The degree of accuracy in the structural model estimates are shown in Figure 13 and presented in Table 8.

Correct Specification, Reflective Indicators. Pair wise comparisons yielded no differences in accuracy of structural model estimates between ML and MCMC for correctly specified models with reflective indicators. For analyses within the context of correctly specified models with reflective indicators, ML and MCMC results were combined. Large effects of sample size ($F(2) = 2125.78$, $p < .001$, partial $\eta^2 = 0.45$), number of items per latent variable ($F(2) = 749.83$, $p < .001$, partial $\eta^2 = 0.22$), and estimation method ($F(2) = 3314.48$, $p < .001$, partial $\eta^2 = 0.56$) were found for accuracy of structural model estimates under conditions of correct model specification and reflective indicators. A moderate sample size × number of items × estimation method interaction effect was also found ($F(8) = 96.11$, $p < .001$, partial $\eta^2 = 0.13$). Under conditions of correctly specified models with reflective indicators, no effects were identified for ML/MCMC estimation, which indicates that these methods performed consistently across all levels of sample size and number of items.

Large effects of sample size and number of items were found for both PLS (sample size: $F(2) = 1538.11$, $p < .001$, partial $\eta^2 = 0.70$; number of items: $F(2) = 134.96$, $p < .001$, partial $\eta^2 = 0.17$) and GSCA (sample size: $F(2) = 860.80$, $p < .001$, partial $\eta^2 =$

0.56; number of items: $F(2) = 835.43$, $p < .001$, partial $\eta^2 = 0.56$), as was a moderate sample size × number of items interaction (PLS: $F(4) = 49.95$, $p < .001$, partial $\eta^2 = 0.13$; GSCA: $F(4) = 183.29$, $p < .001$, partial $\eta^2 = 0.35$). For both PLS and GSCA, accuracy of the structural model estimates increased as the number of items increased, but decreased as sample size increased. ML and MCMC produced more accurate estimates of structural model parameters than PLS and GSCA across all levels of sample size and number of indicators. PLS and GSCA only produced estimates of comparable accuracy to those of ML and MCMC when sample size was small (i.e., $n = 50$) and number of items per latent variable was large (i.e., $i = 7$).

Correct Specification, Formative Indicators. Under conditions of correctly specified models with formative indicators, an effect of estimation method was not found ($F(3) = 5.35$, $p < .001$, partial $\eta^2 < 0.01$), which indicates roughly equivalent performance of the estimation methods within the context of formative indicators and a correctly specified model. Moderate or large effects of sample size and number of items on accuracy of structural model estimates were also not found, which indicates relatively consistent performance of each estimation method across all conditions of sample size and number of indicators when a model is correctly specified and includes only formative latent variable-indicator relationships. It is worth noting that accuracy of GSCA estimates for structural models decreased as the number of items increased, but the magnitude of this decrease was small and the proportion of accurate structural parameter estimates at the largest sample size was greater than 91%. It is also worth noting that accuracy of

structural model estimates for MCMC with $n = 50$ was lower for 5 items than for 3 or 7 items.

Misspecification, Reflective Indicators. Pair wise comparisons yielded no differences in accuracy of structural model estimates between ML and MCMC for misspecified models with reflective indicators. For analyses within the context of misspecified models with reflective indicators, ML and MCMC results were combined. A large effect of sample size ($F(2) = 543.53$, $p < .001$, partial $\eta^2 = 0.17$) was found for accuracy of structural model estimates, with a decrease in accuracy as sample size increased. Under conditions of misspecified models with reflective indicators, a large effect of sample size was observed for ML/MCMC ($F(2) = 734.22$, $p < .001$, partial $\eta^2 = 0.36$), where accuracy of structural model estimated decreases as sample size increases. Similar results were found for PLS ($F(2) = 182.43$, $p < .001$, partial $\eta^2 = 0.21$) and GSCA ($F(2) = 72.45$, $p < .001$, partial $\eta^2 = 0.10$), with accuracy of estimates decreasing as sample size increased. Across all levels of sample size and number of items per latent variable, PLS and GSCA recovered the largest proportion of accurate parameter estimates for the structural model, except in the case of a small sample (i.e., $n = 50$) and large number of items (i.e., $i = 7$), where PLS recovered parameter estimates with the most accuracy but GSCA recovered the least accurate estimates.

Misspecification, Formative Indicators. Moderate and large effects of sample size ($F(2) = 256.40$, $p < .001$, partial $\eta^2 = 0.10$), number of items per latent variable ($F(2) = 1540.30$, $p < .001$, partial $\eta^2 = 0.39$), and estimation method ($F(3) = 1922.47$, $p < .001$, partial $\eta^2 = 0.55$) were found for accuracy of structural model estimates under conditions

of model misspecification with formative indicators. A large sample size × number of items × estimation method interaction effect was also found ($F(12) = 69.74$, $p < .001$, partial $\eta^2 = 0.15$). Under conditions of model misspecification with formative indicators, a large effect of sample size was found for ML ($F(2) = 123.44$, $p < .001$, partial $\eta^2 = 0.19$), PLS ($F(2) = 499.80$, $p < .001$, partial $\eta^2 = 0.43$), and GSCA ($F(2) = 420.34$, $p < .001$, partial $\eta^2 = 0.39$). For each method, accuracy of standard error estimates decreased as sample size increased. A large effect of number of items was found for ML ($F(2) = 545.11$, $p < .001$, partial $\eta^2 = 0.51$), PLS ($F(2) = 123.70$, $p < .001$, partial $\eta^2 = 0.16$), GSCA ($F(2) = 184.65$, $p < .001$, partial $\eta^2 = 0.22$), and MCMC ($F(2) = 1198.34$, $p < .001$, partial $\eta^2 = 0.70$), where accuracy of standard error estimates increased as the number of items increased.

Summary

This study examined the performance of four estimation approaches across 36 experimental conditions for seven key outcomes. The relative performance of the estimation methods was found to be largely dependent on the outcome variable. Table 9 indicates the conditions under which each method performed best for each outcome, compared to the other estimation methods.

Table 9. Summary of top performing estimation methods per experimental condition

		ML	PLS	GSCA	MCMC
Goodness of Fit					
Correct Specification	Reflective	✓			✓
	Formative	✓	✓		
Misspecified	Reflective		✓	✓	
	Formative	✓			✓
Bias of Measurement Model Parameter Estimates					
Correct Specification	Reflective	✓			✓
	Formative	✓	✓		✓
Misspecified	Reflective	✓			✓
	Formative	✓			✓
Bias of Structural Model Parameter Estimates					
Correct Specification	Reflective	✓			✓
	Formative	✓	✓	✓	✓
Misspecified	Reflective		✓	✓	
	Formative	✓			✓
MAD of Standard Error Estimates for Measurement Model					
Correct Specification	Reflective	✓	✓	✓	
	Formative	✓	✓	✓	
Misspecified	Reflective	✓	✓	✓	
	Formative	✓	✓	✓	
MAD of Standard Error Estimates for Structural Model					
Correct Specification	Reflective	✓			✓
	Formative	✓			✓
Misspecified	Reflective	✓			✓
	Formative	✓			✓
Accuracy of Standard Error Estimates for Measurement Model					
Correct Specification	Reflective	✓			✓
	Formative				
Misspecified	Reflective	✓			✓
	Formative				
Accuracy of Standard Error Estimates for Structural Model					
Correct Specification	Reflective	✓			✓
	Formative	✓	✓	✓	✓
Misspecified	Reflective		✓	✓	
	Formative	✓			✓

Goodness of Fit

ML and MCMC consistently produce less biased GOF estimates than either PLS or GSCA when applied to correctly specified models consisting of reflective measurement model relationships and misspecified models consisting of formative measurement model relationships. For correctly specified models with formative measurement model relationships, ML and PLS outperform GSCA and MCMC, with near-zero bias in GOF estimates regardless of sample size or number of items per latent variable. The ability of GSCA and MCMC to recover GOF is influenced by number of items and sample size under conditions of correct specification and formative indicator-latent variable relationships. Specifically, MCMC performs better when the model includes more than three items per latent variable, and GSCA performs better when the model includes fewer than seven items per latent variable and the sample size is small (i.e., $n = 50$). ML and MCMC are outperformed by PLS and GSCA only under conditions of model misspecification with reflective indicator-latent variable relationships.

Bias of Measurement Model Parameter Estimates

ML and MCMC consistently yield the least biased estimates of measurement model parameters across all conditions of model misspecification and type of indicator-latent variable relationships. Under conditions of correct model specification and formative measurement model relationships, PLS performs as well as ML and MCMC, with all three methods producing parameter estimates with near-zero bias. When the model is correctly specified and includes formative relationships in the measurement model, GSCA performs better (relative to itself) when the number of items per latent

variable is small, regardless of sample size. All four estimation methods produce biased parameter estimates for the measurement model when applied to misspecified models with formative indicator-latent variable relationships, though ML and MCMC produce estimates that are less biased than those of PLS or GSCA.

Bias of Structural Model Parameter Estimates

ML and MCMC produce the least biased estimates across all levels of sample size and number of indicators per latent variable under conditions of correct model specification and reflective indicator-latent variable relationships. There is no difference between estimation methods in the bias of the recovered parameter estimates for the structural model under conditions of correct model specification with formative indicator-latent variable relationships in the measurement model. PLS and GSCA produce the least biased estimates across all levels of sample size and number of indicators per latent variable under conditions of model misspecification and reflective indicator-latent variable relationships. ML and MCMC produce the least biased estimates across all levels of sample size and number of indicators per latent variable under conditions of model misspecification and formative indicator-latent variable relationships.

The amount of bias in the parameter estimates decreases as the number of items per latent variable increases for all estimation methods under conditions of misspecified models regardless of the type of indicator-latent variable relationships in the measurement model. Parameter estimates are characterized by less bias for all estimation methods as sample size increases for correctly specified models with formative indicator-latent variable relationships and misspecified models with reflective indicator-latent

variable relationships. Under conditions of model misspecification and formative indicator-latent variable relationships, increased sample size improves the quality (i.e., lowers the bias) of the ML and MCMC parameter estimates.

Mean Differences of Standard Error Estimates for Measurement Models

ML, PLS, and GSCA yield standard error estimates close to the population values (i.e., small MAD) across all levels of sample size and number of indicators per latent variable under all conditions of model specification and type of indicator-latent variable measurement model relationships. MCMC performs well when applied to models with fewer than seven items across all levels of sample size under all model specification and measurement model relationship conditions.

Mean Differences of Standard Error Estimates for Structural Models

Compared to PLS and GSCA, ML and MCMC produce less biased standard error estimates for the structural model across all levels of sample size and number of items per latent variable under conditions of correct model specification (regardless of the type of indicator-latent variable relationship), as well as when the model is misspecified and includes reflective indicator-latent variable relationships in the measurement model. For misspecified models with formative measurement model relationships, ML and MCMC outperform PLS and GSCA across all levels of sample size and number of indicators per latent variable; PLS consistently outperforms GSCA.

Accuracy of Standard Error Estimates for Measurement Models

ML and MCMC produce more accurate estimates for measurement models than PLS and GSCA when the measurement model consists of reflective indicator-latent

variable relationships, regardless of whether the model is specified correctly. When the measurement model consists of formative indicator-latent variable relationships, all methods perform poorly across all levels of sample size and number of indicators per latent variable, except for MCMC. MCMC yields more accurate measurement model parameter estimates when the measurement model includes seven items per latent variable.

Accuracy of Standard Error Estimates for Structural Models

ML and MCMC produced the most accurate estimates of structural model parameters across all levels of sample size and number of items per latent variable under conditions of correct model specification with reflective indicator-latent variable relationships and model misspecification with formative indicator-latent variable relationships. PLS and GSCA performed almost as well as ML and MCMC when applied to data with a small sample size (n = 50) and large number of items (i = 7) under conditions of correct model specification and reflective indicators. Under conditions of correct model specification and formative indicator-latent variable relationships, all four estimation methods performed well across all levels of number of items per latent variable, particularly with samples larger than n = 50. Overall, PLS and GSCA outperformed ML and MCMC under conditions of model misspecification with reflective indicator-latent variable relationships. However, GSCA produced the least accurate structural model estimates of any estimate method under these conditions when n = 50 and the measurement model included 7 items per latent variable.

CHAPTER V. DISCUSSION

This study constitutes a first attempt to compare the relative performance of ML, PLS, GSCA, and MCMC simultaneously under complex data conditions. The overarching goal of this study is to understand the effects of sample size, number of items per latent variable, model misspecification, and the nature of the latent variable-indicator relationships on the performance of ML, PLS, GSCA, and MCMC.

Convergence Rate

Despite its commonplace application in SEMs across disciplines, one limitation of the ML approach to model estimation is its difficulty reaching converged solutions with plausible values as models increase in complexity and sample size decreases. This characteristic was observed in the present study, as the success rate of ML convergence increased as sample size increased and number of items per latent variable decreased. Further, ML successfully converged a larger proportion of times for models that were correctly specified and consisted of reflective indicator-latent variable relationships. Researchers have noted that an advantage of the PLS and GSCA approaches is that they consistently converge to produce plausible value estimates (Fornell & Bookstein, 1982; Hanafi, 2007; Henseler, 2012; Hwang & Takane, 2004). The present research provides additional support for these claims, as PLS and GSCA converged to plausible values for all replications across all experimental conditions included in this study.

Hypothesis 1. It was hypothesized that ML would yield an overall lower convergence rate than PLS, GSCA, or MCMC. This hypothesis was partially supported, as the ML convergence rate of 94% was lower than the 100% convergence rates for PLS

and GSCA. However, the convergence rate observed for ML was not lower than that of the MCMC approach, which only converged to plausible values for 87.7% of replications across experimental conditions. The lower convergence rate for ML compared to PLS is consistent with previous research comparing the two methods (e.g., Hulland et al., 2010; Tenenhaus et al., 2005), as is the lower convergence rate for ML compared to GSCA (e.g., Hwang, Malhotra, et al., 2010). The overall pattern of convergence rates for ML is consistent with that reported previously, in that ML has more frequently failed to converge or failed to recover plausible values under conditions of small sample size (n = 50) and model misspecification (Jackson, 2007), as well as formative indicator-latent variable relationships (Boomsma & Hoogland, 2001; Jackson, 2007).

Research Questions

Research Question 1

To what extent does sample size affect the relative ability of the estimation methods to estimate global model fit and accurately recover model parameters and their standard errors?

Across all levels of number of items per latent variable, degree of misspecification, and nature of the latent variable-indicator relationship, sample size had no effect on the ability of the four estimation methods to estimate global model fit (GOF).

Sample size was found to impact the ability of the estimation methods to recover measurement model parameters across all levels of number of items per latent variable, degree of misspecification, and nature of the latent variable-indicator relationship. For all four estimation methods, bias in the measurement model parameter estimates decreased

as sample size increased. Thus, a lack of differences between estimation methods when the sample size was smallest (n = 50) may indicate that the four methods did an equally poor job recovering measurement model parameters with such a small sample. Differences between the four estimation methods emerged as the sample size increased, with ML and MCMC producing less biased measurement model estimates than either PLS or GSCA when the sample was larger than n = 50 (i.e., when n = 300 and when n = 1000). The ability of MCMC to recover unbiased estimates across all levels of sample size is not surprising given that the method does not rely on an assumption of large samples (Lee & Song, 2004; Song & Lee, 2006). The fact that ML outperformed PLS and GSCA when n = 300 is consistent with the work of Reinartz et al. (2009), who found that ML outperformed PLS when applied to samples of n > 250. The superior performance of ML over PLS and GSCA under the large sample size condition (n = 1,000) is also not surprising, given that ML is known to perform well when its assumptions are met (e.g., Bollen, 1989). Across all levels of number of items per latent variable, degree of misspecification, and nature of the latent variable-indicator relationship, sample size did not impact the ability of the estimation methods to recover structural model parameters.

Sample size was found to impact the ability of the estimation methods to recover standard error estimates for both the measurement and structural models. Across all levels of number of items per latent variable, degree of misspecification, and nature of the latent variable-indicator relationship, differences in the performance of the four estimation methods were observed for mean differences of the measurement model estimates, mean differences of the structural model estimates, and the accuracy of the

recovered structural model estimates. Measurement model estimates recovered by ML, PLS, and GSCA were found to improve (i.e., be less biased) as sample size increased. At all levels of sample size, ML was found to outperform the other three methods, and PLS and GSCA were found to perform better than MCMC. For estimates of the structural model parameters, less biased estimates were recovered by all four estimation methods as sample size increased across all levels of number of items per latent variable, degree of misspecification, and nature of the latent variable-indicator relationships.

Within each level of sample size, ML and MCMC produced less biased estimates than either PLS or GSCA. PLS produced less biased estimates for the structural model compared to GSCA when the sample size was smallest, but GSCA outperformed PLS in both of the larger sample size conditions (i.e., $n = 300$ and $n = 1000$).

With regard to accuracy of the estimates recovered for the measurement model both PLS and GSCA produced fewer accurate parameter estimates as sample size increased. Across all levels of sample size, MCMC outperformed all other estimation methods, and ML and GSCA performed similarly well and both yielded a larger proportion of accurate estimates than PLS. This finding is partially consistent with results reported by Hwang, Malhotra, et al. (2010), in that both Hwang et al. and the present study found that GSCA yields more accurate standard error estimates than PLS. However, Hwang et al. also found GSCA to outperform ML, whereas the two methods performed equally well in the present study.

***Hypothesis 2*.** It was hypothesized that ML, PLS, and MCMC would perform better as sample size increased. This hypothesis was partially supported by the findings of

the present study. ML, PLS, and MCMC did recover less biased parameter estimates for the measurement model as sample size increased, as well as less biased standard error estimates for both the measurement and structural models. This trend did not extend to the accuracy of the standard error estimates for the measurement model, however, as accuracy was found to decrease as sample size increased. The performance of these methods with regard to parameter estimates and bias of standard error estimates is consistent with previous research which demonstrates improved performance of the three estimation methods as sample size increases (Hwang, Malhotra, et al., 2010; Ringle et al., 2009).

Hypothesis 3. It was hypothesized that when the sample size was its smallest, ML would produce more biased estimates of parameters and standard errors than any of the other three estimation methods. This hypothesis was not supported by the present study. Although ML did perform least well under the smallest sample size condition relative to its own performance under conditions of larger sample sizes, it was not found to perform worse than PLS, GSCA, or MCMC across all levels of number of items per latent variable, degree of misspecification, and nature of the latent variable-indicator relationships. In fact, ML was found to recover less biased standard error estimates for the measurement model than either PLS or GSCA when sample size was smallest. This finding is in stark contrast to previous work in this area (e.g., Ringle et al., 2009), but may be explained by the characteristics of the data. Specifically, data for each manifest variable included in the present study were generated to reflect a normal distribution. Such distributions are not likely to be observed in applied research endeavors,

particularly when the sample size is small (i.e., $n = 50$). This may indicate that ML is robust to violations of its sample size assumption when its assumption of normality is tenable.

Research Question 2

To what extent does the number of items per latent variable affect the relative ability of the estimation methods to estimate global model fit and accurately recover model parameters and their standard errors?

Across all levels of sample size, degree of misspecification, and nature of the latent variable-indicator relationship, number of items per latent variable had no effect on the ability of the four estimation methods to estimate global model fit (GOF).

Number of items per latent variable was found to impact the ability of the estimation methods to recover measurement model parameters across all levels of sample size, degree of misspecification, and nature of the latent variable-indicator relationship. For the PLS, GSCA, and MCMC approaches, bias in the measurement model parameter estimates decreased as the number of items increased. Within each level of number of indicators per latent variable, ML recovered less biased parameter estimates than both PLS and GSCA. Under the condition of fewest (3) items, ML performed better than the other estimation methods while MCMC performed worst. Under conditions of an increased number of items per latent variable (i.e., 5 and 7 items), MCMC recovered less biased parameter estimates than either PLS or GSCA. With the largest number of items, PLS recovered less biased parameter estimates than GSCA. Number of items per latent

variable was not found to impact bias of recovered parameter estimates for the structural portions of models.

Number of items per latent variable was found to impact the ability of the estimation methods to recover standard error estimates for measurement models, but did not impact standard error estimates recovered for structural models. Across all levels of sample size, degree of misspecification, and nature of the latent variable-indicator relationship, differences in the performance of the four estimation methods were observed for mean differences of the measurement model standard error estimates, and accuracy of the recovered measurement model estimates. Standard error estimates produced for the measurement portion of models were found to become less biased as the number of items increase for ML, PLS, and MCMC estimation approaches. Number of items did not impact GSCA estimates of standard errors. Performance of the four estimation methods was approximately equal for the smaller number of indicators, but as the number of indicators increased, which contributed to additional measurement model complexity, ML and PLS outperformed GSCA, which in turn outperformed MCMC.

Accuracy of recovered standard error estimates for the measurement model was impacted by number of items, but only for MCMC. Specifically, MCMC estimates were found to be more accurate as the number of items increased. ML recovered the most accurate measurement model estimates with the fewest number of items per latent variable (i = 3), but was outperformed by MCMC with more items. Across all levels of number of items per latent variable, ML and MCMC produced more accurate estimates

for the measurement models than either GSCA or PLS. PLS produced the least accurate measurement model estimates across all levels of number of items per latent variable.

Research Question 3

To what extent does model misspecification affect the relative ability of the estimation methods to estimate global model fit and accurately recover model parameters and their standard errors?

Across all levels of sample size, number of items per latent variable, and nature of the latent variable-indicator relationship, model misspecification was found to impact the ability of the four estimation methods to estimate global model fit (GOF). Specifically, ML, PLS, GSCA, and MCMC overestimated GOF for misspecified models and underestimated GOF for models that were specified correctly. Within the context of misspecified models, ML produced the best estimates of GOF, followed by MCMC, PLS, and then GSCA.

Model misspecification was found to impact the ability of the estimation methods to recover measurement model parameters across all levels of sample size, number of items per latent variable, and nature of the latent variable-indicator relationship. Specifically, all four estimation approaches recovered less biased measurement model parameter estimates under conditions of correct model specification as compared to misspecified models. For conditions of correct model specification as well as model misspecification, ML produced the least biased parameter estimates, followed by MCMC, PLS, and then GSCA.

Degree of model misspecification was found to impact the ability of the estimation methods to recover accurate standard error estimates for measurement models, but did not impact structural model estimates. Of the four estimation methods, degree of model misspecification only impacted the performance of GSCA, which produced more accurate estimates for the measurement model for correctly specified models compared to misspecified models. Within the contexts of both correctly specified and misspecified models, MCMC produced the most accurate estimates for the measurement model, followed by ML. This finding is not entirely consistent with previous research, in that ML has been previously reported to outperform PLS and GSCA only under conditions of correct model specification (e.g., Henseler, 2010; Hwang, Malhotra, et al., 2010). Under conditions of correct model specification, GSCA outperformed PLS with regard to accuracy of measurement model estimates; PLS outperformed GSCA under conditions of model misspecification. These findings are consistent with those of Hwang et al. under conditions of correct model specification, but not entirely consistent under conditions of model misspecification. Hwang et al. found PLS to perform equally well to GSCA under conditions of model misspecification; in the present study, PLS was found to outperform GSCA.

Hypothesis 4. It was hypothesized that under conditions of model misspecification, GSCA would produce less biased estimates of standard errors than ML or PLS. Despite the findings of previous research which indicate GSCA recovers better standard error estimates than other methods when a model is misspecified (e.g., Hwang, Malhotra, et al., 2010), the results of the present study do not support this hypothesis. ML

produced less biased and more accurate standard error estimates than GSCA under both levels of model misspecification, and PLS outperformed GSCA with regard to accuracy of measurement model estimates under misspecification conditions.

Hypothesis 5. It was hypothesized that PLS would recover less biased parameter and standard error estimates for the measurement model when the model is correctly specified compared to when the model is misspecified. This hypothesis was supported; this finding is consistent with previous research (i.e., Haenlin & Kaplan, 2004) and might be interpreted as support for the argument that measurement model estimates are not overly influenced by less-than-perfect structural model estimates because the PLS approach identifies its final solution for the measurement model before arriving at its final estimated values for the structural model. This position is further supported by Hwang et al.'s (2010) finding that PLS recovered unbiased estimates for the measurement model and biased estimates for the structural model.

Research Question 4

To what extent does the nature of the latent variable-indicator relationship affect the relative ability of the estimation methods to estimate global model fit and accurately recover model parameters and their standard errors?

Across all levels of sample size, number of items per latent variable, and model misspecification, the nature of the latent variable-indicator relationship was found to impact the ability of the four estimation methods to estimate global model fit (GOF). Specifically, ML, PLS, GSCA, and MCMC overestimated GOF for formative models and underestimated GOF for reflective models. Within the context of formative latent

variable-indicator relationships, ML produced the best estimates of GOF, followed by MCMC, PLS, and then GSCA.

The nature of the latent variable-indicator relationships was found to impact the ability of the estimation methods to recover parameter estimates for measurement models, parameter estimates for structural models, and standard errors for measurement models. The nature of the measurement model relationships only impacted the recovery of measurement model parameters for PLS, which produced less biased parameter estimates for formative models than for reflective models. This finding is consistent with previous research which indicates that PLS performs especially well when applied to formative measurement models (e.g., Fornell & Bookstein, 1982; MacCallum & Browne, 1993). Within the context of reflective relationships, ML produced parameter estimates for the measurement model with the least amount of bias, followed by PLS, MCMC, and then GSCA. Within the context of formative relationships, ML produced parameter estimates for the measurement model with the least amount of bias, followed by MCMC, PLS, and then GSCA. The nature of the measurement model relationships also impacted the recovery of structural model parameters only for ML, which recovered less biased estimates for formative models than reflective models. Although no difference in parameter estimate bias for the structural model was observed for models with reflective measurement models, differences were observed for models with formative measurement models, where ML produced structural model estimates with the least amount of bias, followed by MCMC, PLS, and finally GSCA.

The nature of latent variable-indicator relationships was also found to impact the performance of ML, PLS, and GSCA with regard to bias of standard error estimates recovered for the measurement model, with all three estimation methods performing better within formative models than reflective models. Within the context of reflective models, ML outperformed the other three estimation methods, followed by PLS, GSCA, and MCMC. Within the context of formative models, ML, PLS, and GSCA all outperformed MCMC, which produced the largest amounts of difference between recovered standard error estimates and their true values. The accuracy of estimates for the measurement model was also impacted by the nature of the latent variable-indicator relationships, as is illustrated by the fact that ML, PLS, GSCA, and MCMC produced a higher proportion of accurate estimates under conditions of reflective measurement models than formative models. Within the context of reflective models, MCMC recovered estimates with the highest accuracy, followed by ML, PLS, and then GSCA. Within the context of formative models, MCMC performed best, followed by GSCA, and then ML and PLS. This pattern of performance for ML (i.e., recovery of more accurate parameter estimates for reflective models than formative models) is consistent with previous research indicating that ML performs better under conditions of reflective measurement models than formative measurement models (e.g., Chin, 1998; Jackson, 2001).

General Discussion

This study attempted to replicate and extend previous research by simultaneously evaluating the performance of four approaches to estimating SEMs using models and

methodological approaches not uncommon in the existing simulation-based research which guided the development of this endeavor. The results presented herein do not precisely replicate the findings reported in previous work, but this difference is not entirely surprising. The present study was a first step at simultaneously comparing four methods that had not been considered in this way before, and the resulting design and methodology were not a strict replication of any one research report. Broadly speaking, the results of the present study indicate that ML may be the best choice among the estimation methods, even when the sample size is small. The overall superior performance of ML over PLS, GSCA, and MCMC was not expected. Several characteristics of the present study may serve to explain or partially explain why ML performed so strongly within the context of the present study. Three likely explanations might be the methods used for simulating the data, the strength of the relationships within the measurement model, and the method used to obtain the true (or, empirical) standard error estimates.

The present study consisted of 150 replication data sets for each of 36 experimental conditions. Each replication data set was simulated directly by the software program to have a specific set of population values. An alternative approach to creating the replication data sets would have been to generate a large population of data for each of 12 experimental conditions (number of items × level of misspecification × nature of indicator-latent variable relationships) from which 150 samples could be drawn for each sample size condition. Both approaches rely on the same population values in the generation of the overall sample, but the sampling approach would be expected to result

in less "perfect" data within each replication data set. The estimation methods examined herein might have performed differently with these data sets, as the normality assumption would not be as strong due to the nature of the sampling process.

Another plausible explanation for the strong performance of ML in the present study is that across all experimental conditions, the relationships (i.e., factor loadings) between indicators and latent variables in the measurement models were quite strong. ML is known to perform better when applied to strong and consistent relationships compared to its performance with weak relationships. It is possible that the assumption of large samples is only particularly important to the performance of ML when some or all of its other assumptions are not met.

A third explanation for the high quality of the parameter and standard error estimates yielded by ML in the present study is the method by which the true values for standard errors of parameter estimates were obtained. Although the standard error estimates were obtained through bootstrapping, those bootstrapping results were based on parameter estimates recovered using ML. Utilizing ML to obtain the parameter estimates for the bootstrapped standard error values may have provided the ML estimation method an unfair advantage over other estimation methods. In addition, it is important to note that ML, PLS, and GSCA all used different software to obtain the bootstrap estimates. It is likely that there are small differences between the software packages in the implementation of the bootstrap process. Because PLS and GSCA bootstrap estimates were not recovered using the same software as was used for obtaining either the true values or the ML estimates, it is possible that these methods were at a disadvantage.

Covariance- *vs.* Component- Based Approaches

A primary expectation at the onset of this research was the importance of the distinction between covariance-based estimation methods (i.e., ML) and component-based approaches (i.e., PLS, GSCA). Because covariance-based estimation methods such as ML make more (and more stringent) assumptions about the data and the nature of the relationships between variables than the component-based approaches, it was anticipated that ML would not perform as well as PLS or GSCA when applied to smaller samples. The pattern of results across all conditions clearly indicates a difference between the covariance- and component-based approaches; surprisingly, the overall results favor the covariance-based ML method over the component-based PLS and GSCA approaches. It appears that ML is robust to the experimental conditions employed for this research. Except for the assumption of large samples, no other assumptions of ML were intentionally violated. It is possible that a more important predictor of ML performance in the estimation of SEMs may be violation of another assumption or some combination of assumptions. This would be consistent with Bentler and Yuan (1999), who posited that the application of ML to small samples is not particularly problematic if its normality assumption remains tenable.

Frequentist *vs.* Bayesian Approaches

The distinction between frequentist (i.e., ML, PLS, GSCA) and Bayesian (i.e., MCMC) approaches to estimating SEMs was also expected to be an important design factor. Despite the expected differences in performance between MCMC and the other estimation methods, no formal hypotheses were generated a priori because there was no

research on which to base expectations for relative performance of MCMC, PLS, and GSCA. With regard to the lower rate of convergence observed for MCMC compared to the other estimation methods, the relatively poor performance of MCMC may be due to characteristics of the estimation process other than the data or models used. Specifically, model convergence may have been greater for the MCMC approach had informative priors been provided (whereas the present study relied entirely on non-informative priors), the burn-in phase extended, or seed values specified a priori. With regard to the recovery of parameter and standard error estimates, differences in the overall performance of MCMC relative to the frequentist approaches were not consistent. Specifically, MCMC performed very similarly to ML under many conditions, but not very similarly to PLS or GSCA. One explanation for this inconsistent distinction between the frequentist and Bayesian approaches might be that ML is both a frequentist and covariance-based method, whereas PLS and GSCA are frequentist, component-based methods. Thus, it appears the covariance *vs.* component distinction may be more important than the distinction between frequentist and Bayesian.

Limitations & Future Research

Despite the unique contribution of this study as a simultaneous comparison of ML, PLS, GSCA, and MCMC estimation methods, it is not without limitations. These limitations include the non-convergence rates of ML and MCMC, simplicity of the population and analytic models, and reliance on the GOF index for evaluation of overall model fit across all estimation methods.

One limitation of the present study is that ML and MCMC failed to converge to plausible values under some conditions. Even though it was expected that ML and MCMC might have difficulty converging under some of the experimental conditions selected, the lower rates of non-convergence for these methods resulted in a smaller sample size for these methods compared to PLS and GSCA. It is possible that the results of this study might be different if 150 converged solutions had been obtained for all estimation methods under all conditions. To investigate this possibility, future researchers might generate more replication data sets than needed, estimate the appropriate model(s) for each data set using all estimation methods, omit any replication data set for which all models and methods did not converge to plausible values, and then randomly sample from the remaining data sets to obtain the desired sample size.

A second limitation of the present study is the simplicity of the population and analytic models. The population models used for the present study were relatively simple compared to some models employed by substantive researchers. Specifically, all data were generated as normally distributed representations of their respective variables, but typical data is rarely normally distributed. Both the population and analytic models used for the present study were relatively simple: each latent variable was related to an equal number of indicators in the measurement models, and the structural models included only a minimal number of latent variables and relationships between those latent variables. The simplicity of the models examined were appropriate for the present study, given its uniqueness in comparison of these four estimation methods. Future research, however, should examine the relative performance of these estimation methods when applied to

more complex models (e.g., cross-loadings as part of the analytic model, combination of reflective and formative indicators in the measurement model, misspecification in the structural portion of the model, multiple group analyses, etc.).

The simplicity of the analytic models may have been particularly problematic when considered in the context of PLS estimation. PLS offers two approaches to model estimation: Mode A for reflective indicators and Mode B for formative indicators. In the present study, the formative measurement model conditions were implemented through reflective measurement models equivalent to a formative model (i.e., high factor loadings and near-zero item reliability), For this reason, Mode A was used for PLS estimation of all models under all conditions. Because the models estimated for the formative model conditions technically were reflective models, Mode A is expected to have performed adequately. However, because the reflective models were essentially formative models, it could be that Mode B would have been a more appropriate choice, and the results of PLS estimation presented here may have been tainted by the use of Mode A and might not reflect the results that would be obtained if Mode B were applied instead. Even though it is not expected that the approach used for the present study negatively influenced the estimation process or recovered estimates, it is a question worth empirical investigation.

Finally, this study was limited in that in order to provide a single, consistent index of global model fit across all estimation methods, the researcher relied on the GOF index. While there is no research which indicates the GOF is unsuitable for methods other than PLS (which is was developed for), there is also a dearth of evidence in support of using GOF with estimation methods that are not component-based (i.e., ML, MCMC). Future

research would do well to investigate the performance of the GOF index of model fit when applied to ML and MCMC models relative to the fit indices typical to those approaches.

In addition to future research designed to overcome these limitations, applied researchers might greatly benefit from consideration of the use of some combination of more than one of these estimation methods from an empirical perspective. The purpose of such an investigation would be to determine whether research would be better served by utilizing a combination of estimation methods for parameter recovery, thereby obtaining some combination of estimates from different methods. If possible, such a practice would allow researchers to benefit from the combined strengths of more than one method instead of selecting the one that appears to be the least flawed for their purposes.

Implications and Conclusions

The driving force behind the need for this type of research is to provide a foundation of information on which applied researchers might rely when selecting an estimation method. To this end, it is imperative that the strengths and weaknesses of existing methods be fully explored and better understood, additional methods developed to bridge gaps between existing approaches, and all of this information made accessible to applied researchers. The present study was a first step at exploring and understanding the performance of component-based estimation methods relative to the traditional ML approach. Although the design of the present study was appropriate for investigating the differential performance of these approaches, it is in no way comprehensive. As a result,

the findings are best interpreted as guidance for the development of additional methodological work to extend this research and delve deeper into the issues at hand.

Applied researchers are cautioned to remember that the results presented here are contingent upon the characteristics of the data generated for this study (i.e., normally distributed variables throughout the measurement and structural models) - characteristics uncommon in substantive, "real-world" research endeavors. This study does, however, emphasize the importance of the estimation method on results, which are directly responsible for conclusions that one might draw. It may be concluded, then, that the choice of estimation method should be based on consideration of the design features of the study at hand (e.g., sample size, number of items per latent variable, possibility of model misspecification, type of indicator-latent variable relationships), the characteristics of the data (e.g., whether the data are normally distributed, the measurement scales used, the strength of the relationships between indicators and latent variables), and which outcomes (i.e., goodness of fit, estimates specific to the measurement model, estimates specific to the structural model, parameter estimates, standard error estimates) will be analyzed and interpreted for the purposes of evaluating a study or developing subsequent projects. The results of the present study indicate that ML is generally the most robust of the four estimation methods studied, and should therefore be the approach of choice for applied researchers. It is important to note, however, that such a conclusion is dependent upon the characteristics of the data and models used for the present study, and this recommendation may not generalize to other contexts.

References

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness of fit indices for MLE CFA. *Psychometrika, 49*, 155-173.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*, 411-423.

Asparouhov, T., & Muthén, B. (2010). *Bayesian analysis using Mplus: Technical implementation*. Retrieved December 18, 2012 from http://www.statmodel.com

Babakus, E.,Ferguson, C. E., & Jöreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research, 24*(2), 222-228.

Bagozzi, R. P., & Yi, Y. (1994). Advanced topics in structural equation models. In: Bagozzi, R. P. (Ed.) *Advanced methods of marketing research*. Blackwell, Oxford, pp 1-51.

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

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

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*(2), 305-314.

Boomsma, A. (1982). Robustness of LISREL against small sample sizes in factor analysis models. In K. G. Joreskog & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction* (Part 1), pp. 149-173. Amsterdam: North Holland.

Boomsma, A., & Hoogland, J. J. (2001). The robustness of LISREL modeling revisted. In R. Cudeck, S. du Tiout, & D. Sörbom (Eds.), *Structural equation models: Present and future. A Festschrift in honor of Karl Jöreskog* (pp. 139-168). Chicago: Scientific Software International.

Brown, T. A. (2006). Confirmatory factor analysis for applied research. New York, NY: Guilford Press.

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis, 1*(3), 473-514.

Carlin, B. P., & Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Part B, 57*(3), 473-484.

Cassel, C., Hackl, P., & Westlund, A. (1999). Robustness of partial least squares method of estimating latent variable quality structures. *Journal of Applied Statistics, 26*, 435-446.

Chin, W. W. (1998). Issues and opinion on structural equation modeling. *Management Information Systems Quarterly, 22*(1).

Chin, W. W., & Newsted, P. R. (1999). Structural equation modeling analysis with small samples using partial least squares. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 307-341). Thousand Oaks, CA: Sage.

Chumney, F. L. (2012). *Comparison of maximum likelihood, Bayesian, partial least squares, and generalized structured component analysis methods for estimation of structural equation models with small samples: An exploratory study.* Unpublished master's thesis, University of Nebraska-Lincoln.

Cohen, J. (1988). *Statistical power analyses for the behavioral sciences*. Hillsdale, NJ: Erlbaum Associates.

Coltman, T., Devinney, T. M., Midgley, D. F., & Venaik, S. (2008). Formative versus reflective measurement models: Two applications of formative measurement. *Journal of Business Research, 61,* 1250-1262.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *American Statistical Association, 91*(434), 883-904.

Curran, P. J., West, S. G., & Finch, J. (1996). The robustness of test statistics to non-normality and specification error in confirmatory factor analysis. *Psychological Methods, 1,* 16-29.

Curtis, R. F., & Jackson, E. F. (1962). Multiple indicators in survey research. *American Journal of Sociology, 68,* 195-204.

de Leeuw, J., Young, F. W., & Takane, Y. (1976). Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika, 41*(4), 471-503.

Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research, 61,* 1203-1218.

Diamantopoulos, A. (2011). Incorporating formative measures into covariance-based structural equation models. *MIS Quarterly, 35*(2), 335-358.

Dijkstra, T. (1983). Some comments on maximum likelihood and partial least squares models. *Journal of Econometrics, 22*, 67-90.

Ding, L., Velicer, W. F., & Harlow, L. L. (1995). Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling, 2*(2), 119-144.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8*(3), 430-457.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling, 6*(1), 56-83.

Fornell, C., & Bookstein, F. L. (1982). Two structural equation models: LISREL and PLS applied to consumer exit-voice theory. *Journal of Marketing Research, 19*, 440-452.

Gerbing, D. W., & Anderson, J. C. (1985). The effects of sampling error and model characteristics on parameter estimation for Maximum Likelihood confirmatory factor analysis. *Multivariate Behavioral Research, 20*, 255-271.

Gerbing, D. W., & Anderson, J. C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 40-65). Newbury Park, CA: Sage.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science, 7*(4), 473-511.

Glover, T. A., Nugent, G., Sheridan, S. M., Bovaird, J. A., & Chumney, F. L. (2013). *Investigating Rural Teachers' Professional Development, Instructional Knowledge, and Classroom Practice.* Manuscript in progress

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103*, 265-275.

Haenlein, M., & Kaplan, A. M. (2004). A beginner's guide to partial least squares analysis. *Understanding Statistics, 3*(4), 283-297.

Hair, J. F., Sarstedt, M., Ringle, C. M., & Mena, J. A. (2012). As assessment on the use of partial least squares structural equation modeling in marketing research. *Journal of the Academy of Marketing Sciece, 40*, 414-433.

Hanafi, M. (2007). PLS path modeling: Computation of latent variables with the estimation mode B. *Computational Statistics, 22*, 275-292.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57*, 97-109.

Henseler, J. (2010). On the convergence of the partial least squares path modeling algorithm. *Computational Statistics, 25*, 107-120.

Henseler, J. (2012). Why generalized structured component analysis is not universally preferable to structural equation modeling. *Journal of the Academy of Marketing Science, 40*, 402-413.

Henseler, J., & Chin, W. W. (2010). A comparison of approaches for the analysis of interaction effects between latent variables using partial least squares path modeling. *Structural Equation Modeling, 17*, 82-109.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research, 26*(3), 329-367.

Hox, J. J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling, 8*(2), 157-174.

Hoyle, R. H. (2000). Confirmatory factor analysis. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling*. New York: Academic Press.

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

Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112*(2), 351-362.

Hulland, J., Ryan, M. J., & Rayner, R. K. (2010). Modeling customer satisfaction: A comparative performance evaluation of covariance structure analysis versus partial least squares. In V. E. Vinzi, W. W. Chin, J. Henseler, & H. Wang (Eds.), *Handbook of Partial Least Squares*. Berlin: Springer-Verlag.

Hutchinson, S. R., & Bandalos, D. L. (1997). A guide to Monte Carlo simulation research for applied researhcers. *Journal of Vocational Education Research, 22*(4).

Hwang, H. (2009). Regularized generalized structured component analysis. *Psychometrika, 74*(3), 517-530.

Hwang, H., DeSarbo, W. S., & Takane, Y. (2007). Fuzzy clusterwise generalized structured component analysis. *Psychometrika, 72*(2), 181-198.

Hwang, H., Ho, M. H. R., & Lee, J. (2010). Generalized structured component analysis with latent interactions. *Psychometrika, 75*, 228-242.

Hwang, H., Malhotra, N. K., Kim, Y., Tomiuk, & Hong (2010). A comparative study on parameter recovery of three approaches to structural equation modeling. *Journal of Marketing Research, 47*, 699-712.

Hwang, H., & Takane, Y. (2004). Generalized structured component analysis. *Psychometrika, 69*(1), 81-99.

Jackson, D. L. (2001). Sample size and the number of parameter estimates in maximum likelihood confirmatory factor analysis: A Monte Carlo investigation. *Structural Equation Modeling, 8*, 205-223.

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N:q hypothesis. *Structural Equation Modeling, 10*(1), 128-141.

Jackson, D. L. (2007). The effect of the number of observations per parameter in misspecified confirmatory factor analytic models. *Structural Equation Modeling, 14*(1), 48-76.

Jöreskog, K. G., & Wold, H. The ML and PLS techniques for modeling with latent variables: Historical and comparative aspects. In H. Wold & K. Jöreskog (Eds.), *Systems under indirect observation: Causality, structure, prediction II,* (pp. 263-270). Amsterdam: North-Holland.

Kankaraš, M., Vermunt, J. K., & Moors, G. (2011). Measurement equivalence of ordinal items: A comparison of factor analytic, item response theory, and latent class approaches. *Sociological Methods & Research, 40*(2), 279-310.

Kline, R. B. (2011). Principles and practice of structural equation modeling. New York, NY: Guilford.

Lee, S. Y., & Song, X. Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research, 39,* 653-686.

Lee, S. Y., Song, X. Y., & Lee, J. C. K. (2003). Maximum likelihood estimation of nonlinear structural equation models with ignorable missing data. *Journal of Educational and Behavioral Statistics, 28*(2), 111-134.

Lee, S. Y., & Tang, N. S. (2006). Bayesian analysis of nonlinear structural equation mdoels with nonignorable missing data. *Psychometrika, 71*(3), 541-564.

Lee, S. Y., & Xia, Y. M. (2008). A robust Bayesian approach for structural equation models with missing data. *Psychometrika, 73*(3), 343-364.

Lee, S. Y., & Zhu, H. T. (2002). Maximum likelihood estimation of nonlinear structural equation models. *Psychometrika, 67*(2), 189-210.

Lohmöller, J. –B. (1989). *Latent variable path modeling with partial least squares.* Heidelberg, Germany: Physica Verlag.

Lynch, S. M. (2007). Introduction to applied Bayesian statistics and estimation for social scientists. New York: Springer.

Maas, C. J. M., & Hox, J. J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis, 46*, 427-440.

MacCallum, R. C., & Browne, M. W. (1993) The use of causal indicators in covariance structure models: Some practical issues. *Psychological Bulletin, 114*, 533-541.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130-149.

Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33*(2), 181-220.

McDonald, R. P. (1996). Path analysis with composite variables. *Multivariate Behavioral Research, 31*(2), 239-270.

Meyn, S. P., & Tweedie, R. L. (1993). *Markov chain and stochastic stability.* London: Springer-Verlag.

Monecke, A., & Leisch, F. (2012). semPLS: Structural equation modeling using partial least squares. *Journal of Statistical Software, 48*(3), 1-32.

Muthén, B. O. (2010). *Bayesian analysis in Mplus: A brief introduction.* Retrieved December 12, 2012 from http://www.statmodel.com

Muthén, B., & Asparouhov, T. (2011). Bayesian SEM: A more flexible representation of substantive theory. Forthcoming in *Psychological Methods*.

Muthén, L. K., & Muthén, B. O. (1998-2010). M*plus* User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9*, 599-620.

Nevitt, J., & Hancock, G. R. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behavioral Research, 39*(3), 439-478.

Nunnally, J. C. (1967). *Psychometric theory.* New York: McGraw-Hill.

Olsson, U. H., Foss, T., & Breivik, E. (2004). Two equivalent discrepancy functions for maximum likelihood estimation: Do their test statistics follow a non-central chi-square distribution under model misspecification? *Sociological Methods & Research, 32*(4), 453-500.

Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000) The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling, 7*(4), 557-595.

Olsson, U. H., Troye, S. V., & Howell, R. D. (1999). Theoretical fit and empirical fit: The performance of maximum likelihood versus generalized least squares estimation in structural equation models. *Multivariate Behavioral Research, 34*(1), 31-58.

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

Petter, S., Straub, D., & Rai, A. (2007). Specifying formative constructs in information systems research. *MIS Quarterly, 31*(4), 623-656.

R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL *http://www.R-project.org.*

Reinartz, W. J., Haenlein, M., & Henseler, J. (2009). An empirical comparison of the efficacy of covariance-based and variance-based SEM. *International Journal of Market Research, 26*(4), 332-344.

Ringle, C. M., Götz, O., Wetzels, M., & Wilson, B. (2009). On the use of formative measurement specifications in structural equation modeling: A Monte Carlo simulation study to compare covariance-based and partial least squares model estimation methodologies. In *METEOR Research Memoranda (RM/09/014)*: Maastricht University.

Roy, S., Tarafdar, M., Ragu-Nathan, T. S., & Marsillac, E. (2012). The effects of misspecification of reflective and formative constructs in operations and manufacturing management research. *The Electronic Journal of Business Research Methods, 10*(1), 34-52.

Sharma, S., Durvasula, S., & Dillon, W. R. (1989). Some results on the behavior of alternate covariance structure estimation procedures in the presence of non-normal data. *Journal of Marketing Research, 26,* 214-221.

Song, X. Y., & Lee, S. Y. (2002). Analysis of structural equation model with ignorable missing continuous and polytomous data. *Psychometrika, 67*(2), 261-288.

Song, X. Y., & Lee, S. Y. (2005). Maximum likelihood analysis of nonlinear structural equation models with dichotomous variables. *Multivariate Behavioral Research, 40*(2), 151-177.

Song, X. Y., & Lee, S. Y. (2006). Model comparison of generalized linear mixed models. *Statistics in Medicine, 25,* 1685-1698.

Song, X. Y., Lee, S. Y., & Hser, Y. I. (2008). A two-level structural equation model approach for analyzing multivariate longitudinal responses. *Statistics in Medicine, 27,* 3017-3041.

Srinivasan, V., & Mason, C. H. (1986). Nonlinear least squares estimation of new product diffusion models. *Marketing Science 5,* 169-178.

Tanaka, J. S. (1987). "How big is big enough?": Sample size and goodness of fit in structural equation models with latent variables. *Child Development, 58*(1), 134-146.

Temme, D., Kreis, H., & Hildebrandt, L. (2006). *PLS path modeling – A software review.* SFB Discussion Paper 2006-084, Institute of Marketing, Humboldt-Universität zu Berlin, Germany.

Tenenhaus, M. (2008). Component-based structural equation modelling. *Total Quality Management, 19*, 871-886.

Tenenhaus, M., Vinzi, V. E., Chatelin, Y. M., & Lauro, C. (2005). PLS path modeling. *Computational Statistics & Data Analysis, 48*, 159-205.

Tomás, J. M., Hontangas, P. M., & Oliver, A. (2000). Linear confirmatory factor models to evaluate multitrait-multimethod matrices: The effects of number of indicators and correlation among methods. *Multivariate Behavioral Research, 35*(4), 469-499.

Tomás, J. M., Hontangas, P. M., & Oliver, A. (2013). Linear confirmatory factor models to evaluate multitrait-multimethod matrices: The effects of number of indicators and correlation among methods. *Multivariate Behavioral Research, 35*(4), 469-499.

Velicer, W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods, 3*(2), 231-251.

Vinzi, V. E., Trinchera, L., & Amato, S. (2010). PLS path modeling: From foundation to recent developments and open issues for model assessment and improvement. In V. E. Vinzi et al. (Eds.), *Handbook of Partial Least Squares*. Springer-Verlag: Berlin.

Wilcox, J. B., Howell, R. D., & Breivik, E. (2008). Questions about formative measurement. *Journal of Business Research, 61,* 1219-1228.

Wold, H. (1975). Path models with latent variables: The NIPALS approach. In H. M. Blalock, A. Aganbegian, F. M. Borodkin, R. Boudon, & V. Capecchi (Eds.), *Quantitative sociology: International perspectives on mathematical and statistical modeling* (pp. 307-357). New York: Academic.

Wold, H. (1982). Soft modeling: The basic design and some extensions. In H. Wold & K. Jöreskog (Eds.), *Systems under indirect observation: Causality, structure, prediction II,* (pp. 589-591). Amsterdam: North-Holland.

Zhu, B., Walter, S. D., Rosenbaum, P. L., Russell, D. J., & Raina, P. (2006). Structural equation and log-linear modeling: A comparison of methods in the analysis of a study on caregivers' health. *BMC Medical Research Methodology, 6,* 49-62.

APPENDIX A: RESULTS OF FIVE-FACTOR MULTIVARIATE ANALYSIS

Table A.1. Results of multivariate tests for five-factor MANOVA

Effect	Wilks' Λ	df	F
Intercept	.005	7,20460	586426.91
Sample Size (n)	.225	14,40920	3233.71‡
Number of Items per Latent Variable (items)	.002	14,40920	68787.61‡
Degree of Misspecification (spec)	.067	7,20460	40468.22‡
Type of Measurement Model Relationships (iLV)	.084	7,20460	31806.49‡
Estimation Method (em)	.001	21,58751	29280.84‡
n×items	.345	28,73771	903.40‡
n×spec	.716	14,40920	531.37‡
n×iLV	.437	14,40920	1499.90‡
n×em	.440	42,95969	437.74‡
items×spec	.233	14,40920	3130.97‡
items×iLV	.049	14,40920	10309.85‡
items×em	.001	42,95969	8771.86‡
spec×iLV	.118	7,20460	21822.98‡
spec×em	.670	21,58751	418.33‡
iLV×em	.027	21,58751	7004.67‡
n×items×spec	.537	28,73771	496.60‡
n×items×iLV	.360	28,73771	862.07‡
n×items×em	.433	84,125317	218.69†
n×spec×iLV	.746	14,40920	461.23‡
n×spec×em	.715	42,95969	169.53
n×iLV×em	.437	42,95969	441.26‡
items×spec×iLV	.302	14,40920	2392.03‡
items×spec×em	.658	42,95969	213.09†
items×iLV×em	.022	42,95969	2847.35‡
spec×iLV×em	.491	21,58751	785.52‡
n×items×spec×iLV	.522	28,73771	520.27‡
n×items×spec×em	.806	84,125317	53.55
n×items×iLV×em	.501	84,125317	178.19†
n×spec×iLV×em	.754	42,95969	141.58
items×spec×iLV×em	.751	42,95969	143.78
n×items×spec×iLV×em	.646	84,125317	110.43†

Notes: all $p < 0.001$; † partial $\eta^2 > 0.06$; ‡ partial $\eta^2 > 0.13$

Table A.2. Tests of between-subjects effects for Goodness of Fit estimates

Effect	df	F	partial η^2	Effect	df	F	partial η^2
Intercept	1	1173.09	.054	n×items×spec	4	17.55	.003
n	2	36.85	.004	n×items×iLV	4	14.62	.003
items	2	1890.25	.156	n×items×em	12	16.63	.010
spec	1	86023.36	.808	n×spec×iLV	2	9.55	.001
iLV	1	28409.56	.581	n×spec×m	6	13.06	.004
em	3	68.73	.010	n×iLV×em	6	21.84	.006
n×items	4	27.44	.005	items×spec×iLV	2	272.16	.026
n×spec	2	17.12	.002	items×spec×em	6	10.35	.003
n×iLV	2	12.32	.001	items×iLV×em	6	39.90	.012
n×em	6	16.87	.005	spec×iLV×em	3	718.83	.095
items×spec	2	2702.19	.209	n×items×spec×iLV	4	12.39	.002
items×iLV	2	819.60	.074	n×items×spec×em	12	13.41	.008
items×em	6	41.97	.012	n×items×iLV×em	12	12.33	.007
spec×iLV	1	7346.35	.264	n×spec×iLV×em	6	12.49	.004
spec×em	3	769.61	.101	items×spec×iLV×em	6	27.92	.008
iLV×em	3	1296.56	.160	n×items×spec×iLV×em	12	9.01	.005

Notes: all $p < 0.001$; n = Sample Size; items = Number of Items per Latent Variable; spec = Degree of Misspecification; iLV = Type of Measurement Model Relationships; em = Estimation Method

Table A.3. Tests of between-subjects effects for bias of measurement model parameter estimates

Effect	df	F	partial η^2	Effect	df	F	partial η^2
Intercept	1	800932.69	.975	n×items×spec	4	241.82	.045
n	2	3050.40	.230	n×items×iLV	4	402.05	.073
items	2	25353.97	.712	n×items×em	12	337.25	.165
spec	1	166281.82	.890	n×spec×iLV	2	592.30	.055
iLV	1	9599.88	.319	n×spec×m	6	261.07	.071
em	3	20046.30	.746	n×iLV×em	6	903.06	.209
n×items	4	313.90	.058	items×spec×iLV	2	10561.63	.508
n×spec	2	329.77	.031	items×spec×em	6	866.66	.203
n×iLV	2	3358.78	.247	items×iLV×em	6	1397.32	.291
n×em	6	1040.21	.234	spec×iLV×em	3	543.84	.074
items×spec	2	22835.47	.691	n×items×spec×iLV	4	382.97	.070
items×iLV	2	92.25	.009	n×items×spec×em	12	245.85	.126
items×em	6	3181.93	.483	n×items×iLV×em	12	370.36	.178
spec×iLV	1	80734.81	.798	n×spec×iLV×em	6	486.51	.125
spec×em	3	434.75	.060	items×spec×iLV×em	6	286.09	.077
iLV×em	3	2507.73	.269	n×items×spec×iLV×em	12	558.37	.247

Notes: all $p < 0.001$ except where noted; *$p < 0.05$; **$p < 0.01$; n = Sample Size; items = Number of Items per Latent Variable; spec = Degree of Misspecification; iLV = Type of Measurement Model Relationships; em = Estimation Method

Table A.4. Tests of between-subjects effects for bias of structural model parameter estimates

Effect	df	F	partial η^2	Effect	df	F	partial η^2
Intercept	1	71170.46	.777	n×items×spec	4	10.40	.002
n	2	1766.39	.147	n×items×iLV	4	8.97	.002
items	2	2381.45	.189	n×items×em	12	10.29	.006
spec	1	2635.13	.114	n×spec×iLV	2	37.04	.004
iLV	1	1.83 ‖	.000	n×spec×m	6	14.21	.004
em	3	1051.43	.134	n×iLV×em	6	12.08	.004
n×items	4	15.44	.003	items×spec×iLV	2	1345.34	.116
n×spec	2	131.30	.013	items×spec×em	6	114.16	.032
n×iLV	2	.31 ‖	.000	items×iLV×em	6	82.29	.024
n×em	6	72.59	.021	spec×iLV×em	3	2363.04	.257
items×spec	2	415.60	.039	n×items×spec×iLV	4	6.39	.001
items×iLV	2	46.13	.004	n×items×spec×em	12	15.84	.009
items×em	6	35.98	.010	n×items×iLV×em	12	8.62	.005
spec×iLV	1	5279.58	.205	n×spec×iLV×em	6	77.95	.022
spec×em	3	166.04	.024	items×spec×iLV×em	6	55.53	.016
iLV×em	3	540.98	.073	n×items×spec×iLV×em	12	4.35	.003

Notes: all $p < 0.001$ except where noted; ‖$p > 0.05$; n = Sample Size; items = Number of Items per Latent Variable; spec = Degree of Misspecification; iLV = Type of Measurement Model Relationships; em = Estimation Method

Table A.5. Tests of between-subjects effects of MAD for measurement model standard error estimates

Effect	df	F	partial η^2	Effect	df	F	partial η^2
Intercept	1	1670188.52	.988	n×items×spec	4	2688.79	.344
n	2	8329.82	.449	n×items×iLV	4	4519.97	.469
items	2	1461345.13	.993	n×items×em	12	617.87	.266
spec	1	5968.10	.226	n×spec×iLV	2	1418.45	.122
iLV	1	3965.09	.162	n×spec×m	6	8.57	.003
em	3	664127.05	.990	n×iLV×em	6	507.93	.130
n×items	4	5005.87	.495	items×spec×iLV	2	2240.22	.180
n×spec	2	2282.55	.182	items×spec×em	6	22.30	.006
n×iLV	2	6247.36	.379	items×iLV×em	6	59272.27	.946
n×em	6	696.66	.170	spec×iLV×em	3	83.09	.012
items×spec	2	5000.21	.328	n×items×spec×iLV	4	2374.74	.317
items×iLV	2	103992.48	.910	n×items×spec×em	12	8.45	.005
items×em	6	1424279.72	.998	n×items×iLV×em	12	498.28	.226
spec×iLV	1	4855.35	.192	n×spec×iLV×em	6	10.66	.003
spec×em	3	18.78	.003	items×spec×iLV×em	6	129.40	.037
iLV×em	3	34366.03	.834	n×items×spec×iLV×em	12	9.35	.005

Notes: all $p < 0.001$; n = Sample Size; items = Number of Items per Latent Variable; spec = Degree of Misspecification; iLV = Type of Measurement Model Relationships; em = Estimation Method

Table A.6. Tests of between-subjects effects of MAD for structural model standard error estimates

Effect	df	F	partial η^2	Effect	df	F	partial η^2
Intercept	1	27056.69	.569	n×items×spec	4	62.78	.012
n	2	7278.95	.416	n×items×iLV	4	18.03	.004
items	2	469.48	.044	n×items×em	12	150.35	.081
spec	1	1670.08	.075	n×spec×iLV	2	4.33 *	.000
iLV	1	2174.70	.096	n×spec×m	6	7.76	.002
em	3	1226.71	.152	n×iLV×em	6	44.03	.013
n×items	4	39.83	.008	items×spec×iLV	2	2.77 ‖	.000
n×spec	2	163.11	.016	items×spec×em	6	36.31	.011
n×iLV	2	633.57	.058	items×iLV×em	6	53.21	.015
n×em	6	368.59	.098	spec×iLV×em	3	95.84	.014
items×spec	2	581.42	.054	n×items×spec×iLV	4	3.23 *	.001
items×iLV	2	72.70	.007	n×items×spec×em	12	2.78 **	.002
items×em	6	200.91	.056	n×items×iLV×em	12	46.86	.027
spec×iLV	1	28.90	.001	n×spec×iLV×em	6	16.18	.005
spec×em	3	50.53	.007	items×spec×iLV×em	6	27.10	.008
iLV×em	3	63.66	.009	n×items×spec×iLV×em	12	7.28	.004

Notes: all $p < 0.001$ except where noted; *$p < 0.05$; **$p < 0.01$; ‖$p > 0.05$; n = Sample Size; items = Number of Items per Latent Variable; spec = Degree of Misspecification; iLV = Type of Measurement Model Relationships; em = Estimation Method

Table A.7. Tests of between-subjects effects for accuracy of measurement model estimates

Effect	df	F	partial η^2	Effect	df	F	partial η^2
Intercept	1	271032.27	.930	n×items×spec	4	36.90	.007
n	2	7260.94	.415	n×items×iLV	4	235.06	.044
items	2	10279.73	.501	n×items×em	12	172.76	.092
spec	1	2661.46	.115	n×spec×iLV	2	6.69 **	.001
iLV	1	94071.44	.821	n×spec×m	6	594.54	.148
em	3	17379.03	.718	n×iLV×em	6	1454.34	.299
n×items	4	418.26	.076	items×spec×iLV	2	103.45	.010
n×spec	2	76.31	.007	items×spec×em	6	64.38	.019
n×iLV	2	903.93	.081	items×iLV×em	6	4338.88	.560
n×em	6	1674.34	.329	spec×iLV×em	3	929.22	.120
items×spec	2	187.52	.018	n×items×spec×iLV	4	47.39	.009
items×iLV	2	1094.62	.097	n×items×spec×em	12	35.07	.020
items×em	6	3767.15	.525	n×items×iLV×em	12	125.42	.068
spec×iLV	1	1338.31	.061	n×spec×iLV×em	6	142.04	.040
spec×em	3	475.19	.065	items×spec×iLV×em	6	383.47	.101
iLV×em	3	14326.05	.677	n×items×spec×iLV×em	12	29.46	.017

Notes: all $p < 0.001$ except where noted; **$p < 0.01$; n = Sample Size; items = Number of Items per Latent Variable; spec = Degree of Misspecification; iLV = Type of Measurement Model Relationships; em = Estimation Method

Table A.8. Tests of between-subjects effects for accuracy of structural model estimates

Effect	df	F	partial η^2	Effect	df	F	partial η^2
Intercept	1	138991.75	.872	n×items×spec	4	21.42	.004
n	2	1098.17	.097	n×items×iLV	4	64.48	.012
items	2	1080.61	.096	n×items×em	12	41.44	.024
spec	1	4956.30	.195	n×spec×iLV	2	199.18	.019
iLV	1	8.12 **	.000	n×spec×m	6	139.43	.039
em	3	1153.60	.145	n×iLV×em	6	29.53	.009
n×items	4	121.29	.023	items×spec×iLV	2	747.66	.068
n×spec	2	202.84	.019	items×spec×em	6	165.74	.046
n×iLV	2	296.76	.028	items×iLV×em	6	140.54	.040
n×em	6	96.79	.028	spec×iLV×em	3	2888.76	.297
items×spec	2	262.09	.025	n×items×spec×iLV	4	2.78 *	.001
items×iLV	2	54.36	.005	n×items×spec×em	12	20.19	.012
items×em	6	34.29	.010	n×items×iLV×em	12	99.15	.055
spec×iLV	1	3531.68	.147	n×spec×iLV×em	6	242.72	.066
spec×em	3	79.14	.011	items×spec×iLV×em	6	90.74	.026
iLV×em	3	247.87	.035	n×items×spec×iLV×em	12	40.65	.023

Notes: all $p < 0.001$ except where noted; *$p < 0.05$; **$p < 0.01$; n = Sample Size; items = Number of Items per Latent Variable; spec = Degree of Misspecification; iLV = Type of Measurement Model Relationships; em = Estimation Method