

Comparing Bayesian parametric and semiparametric estimation of nonlinear relationships in structural equation models with ordinal data

By

© 2018

Lu Qin

M.S., California State University, San Bernardino, 2011

B.A., Sichuan University, 2008

Submitted to the graduate degree program in Educational Psychology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chair: Jonathan Templin

Brue Frey

David Hansen

Lesa Hoffman

Paul Johnson

Date Defended: Dec 4, 2018

The dissertation committee for Lu Qin certifies that this is the approved
version of the following dissertation:

**Comparing Bayesian parametric and semiparametric estimation
of nonlinear relationships in structural equation models with
ordinal data**

Chair: Jonathan Templin

Date Approved: Dec 4, 2018

Abstract

The Bayesian parametric and semiparametric approaches are compared to recover the polynomial and nonpolynomial relationships among latent factors in the structural equation model (SEM). In earlier studies, the semiparametric approach has been demonstrated to be a more advanced approach to estimate the nonnormally distributed densities. However, its performance in recovering nonlinearity among factors has not been widely studied. The objectives of this dissertation are (1) to compare the recovery performances between the parametric and semiparametric approaches in capturing the polynomial and nonpolynomial relationships among latent factors in the structural model and (2) to investigate the recovery performance of the semiparametric approach in capturing the nonpolynomial relationships when the polynomial function is misspecified. The Bayesian semiparametric approach is applied using the truncated Dirichlet process with a stick-breaking prior to track the nonlinearity under different combinations of nonlinear functions (e.g., exponential, logarithmic, and sine) in the simulation study.

Several important results were revealed. First, in study 1, both the parametric and semiparametric approaches achieved good convergence rates under the exponential and sine conditions. The polynomial conditions had greater difficulty in convergence due to the quadratic and interaction effects. Second, regarding the nonlinearity recoveries, the parametric approach performed similarly to the semiparametric approach at large truncation levels (200) in recovering the polynomial nonlinearity. The semiparametric approach had better recovery of nonpolynomial nonlinearity than the parametric approach. Third, in study 2, the semiparametric approach had a fairly good convergence rate at truncation level 5 under the exponential and sine conditions. Fourth, the semiparametric approach barely recovered the nonpolynomial

nonlinearity with a misspecified polynomial function. A large truncation level did not improve the recovery performance when a nonlinear function is incorrectly presumed.

The results implied that when latent factors or data is normally distributed, parametric approach is sufficient to provide an accurate recovery of nonlinear relationships among latent factors.

However, when latent factors or data is non-normally distributed, the semiparametric approach provides more accurate estimations and a higher accuracy in capturing nonlinear relationships among latent factors. Considering the capacity of computer memory and running time, a small truncation level is suggested to capture the polynomial and nonpolynomial nonlinearity.

Acknowledgements

I want to convey my highest appreciation to my advisors, Dr. Jonathan Templin and Dr. William Skorupski, for their mentorship and guidance through my entire doctoral program. In particular, I want to give many thanks to my advisor, Dr. Jonathan Templin. Thank you for supporting me in choosing such a challenging topic and providing me with endless support from the beginning to the end. I am incredibly appreciative of your patience in walking me through the entire process of my dissertation, including choosing the research topic, correcting codes and drafts, and inspecting typos and errors. Additionally, I am extremely grateful that you have given me different opportunities to learn about myself, along with your constant encouragement.

Many thanks to Dr. Lesa Hoffman, for being such as a great teacher and leading me to the longitudinal and multilevel world. Thank you for being so patient and supportive whenever I ask silly questions and share crazy ideas; thank you for your useful feedback and suggestions in my proposal meeting, which have made this dissertation great; and thank you for sharing your experiences and advice on my career selection.

Thank you to Dr. William Skorupski for opening the door to the Bayesian and simulation world for me. Thank you to Dr. Paul Johnson for providing support in code development. Thank you to Dr. David Hansen for the suggestions in the proposal meeting. Thank you to Dr. Bruce Frey for being supportive at the defense meeting and providing suggestions at the proposal meeting.

I could not have completed the dissertation without the generous help of many people. Thank you, Dr. Qianqian Pan, for sharing information about and teaching me how to use Amazon AWS so that my simulation could be completed as quickly as possible. Thank you to

Dr. Zhehan Jiang for helping me overcome difficulties and for encouraging me through the entire length of my doctoral studies.

Finally, thank you to my family, Zuotao, Amber, Chao, and Jiushu, for your constant consideration, unconditional love, understanding, and encouragement.

Table of Contents

Chapter 1: Introduction	1
Background	1
Statement of Problems	5
Purpose of the Study.....	8
Research Hypotheses	8
Research Questions	9
Plan of the Study	10
Chapter 2: Literature Review	11
Parametric Approaches in Estimating Nonlinear Relations.....	12
Semiparametric Approaches in Estimating Nonnormal Density	13
Semiparametric Approaches in Estimating Nonlinearity	14
Semiparametric SEM (SSEM)	15
Measurement model.....	15
Structural model.....	18
Semiparametric Bayesian Estimation	21
Chapter 3: Method	23
Data Generation.....	23
Generating the measurement model.....	23
Generating the structural model.....	25
Data Analysis	25
Study 1	26
Parametric Bayesian approach	26
Priors distribution.....	26
Posterior distribution.....	27
Semiparametric Bayesian approach.	28
Prior distribution	28
Posterior distribution.....	29
Outcomes.	29
Study 2	32
Outcomes	32

Chapter 4: Results	34
Study 1	34
Nonconvergence rates.	34
Recovery rate	35
Polynomial nonlinear function	35
Exponential nonlinear function	42
Sine nonlinear function.....	47
Study 2	53
Nonconvergence rate.....	53
Recovery rate	54
Chapter 5: Discussion	56
Performance of the Parametric and Semiparametric SEM	59
Model convergence	59
Nonlinearity recoveries	59
Conclusion and Recommendations	60
Contributions and Limitations.....	61
References	63

List of Figures

Figure 1: Polynomial nonlinear relationships between exogeneous latent factors ($\theta Xp, \theta Mp$) and endogenous latent factor (θYp).	3
Figure 2: Exponential nonlinear relationships between exogeneous latent factors ($\theta Xp, \theta Mp$) and endogenous latent factor (θYp).	4
Figure 3: Sinusoidal nonlinear relationships between exogeneous latent factors ($\theta Xp, \theta Mp$) and endogenous latent factor (θYp).	5
Figure 4: Fitting polynomial nonlinear relations in the nonpolynomial nonlinear relations.....	7
Figure 5: The $\theta M - \theta X$ and $\theta Y - \theta M\theta X$ polynomial curves estimated by the parametric and the semiparametric approach.....	40
Figure 6. The mean range of differences in the $\theta M - \theta X$ and $\theta Y - \theta M\theta X$ polynomial curves...	41
Figure 7: The $\theta M - \theta X$ and $\theta Y - \theta M\theta X$ exponential curves estimated by the parametric and semiparametric approaches.....	46
Figure 8: The mean range of differences in the $\theta M - \theta X$ and $\theta Y - \theta M\theta X$ exponential curves.	47
Figure 9: The $\theta M - \theta X$ and $\theta Y - \theta M\theta X$ sine curves estimated by the parametric and semiparametric approaches.....	52
Figure 10: The mean range of differences in the $\theta M - \theta X$ and $\theta Y - \theta M\theta X$ sine curves	53
Figure 11: The exponential and sine curves estimated by the semiparametric approach with the polynomial nonlinear function.....	55

List of Tables

Table 1: Testing the Quadratic Effect within Nonpolynomial Nonlinearity	7
Table 2: A 3 x 4 Simulation Design	23
Table 3: Mean Nonconvergence Rate across 100 Replications	34
Table 4: Mean Nonconvergence Rate across 50 Replications	53

List of Equations

(1).....	2
(2).....	2
(3).....	3
(4).....	3
(5).....	4
(6).....	4
(7).....	4
(8).....	5
(9).....	16
(10).....	16
(11).....	17
(12).....	17
(13).....	17
(14).....	17
(15).....	18
(16).....	18
(17).....	18
(18).....	18
(19).....	19
(20).....	19
(21).....	19
(22).....	19
(23).....	19
(24).....	20
(25).....	20
(26).....	20
(27).....	20
(28).....	20
(29).....	21
(30).....	21
(31).....	26

(32)26
(33)27
(34)27
(35)27
(36)27
(37)28
(38)28
(39)28
(40)29
(41)29
(42)29
(43)30
(44)30
(45)30
(46)30
(47)31
(48)31
(49)31
(50)32
(51)32
(52)32
(53)32
(54)33
(55)33

Chapter 1: Introduction

This study applies a Bayesian semiparametric approach to account for nonlinearity in both the measurement and structural models. The objective of this research is to explore whether the Bayesian semiparametric approach helps recover the true nonlinear relationships in latent structural models without the limitations of distributional assumptions on endogenous latent factors and measurement error.

Background

Nonlinearity in structural equation models (SEMs) can occur in the measurement model and/or the structural model (Kenny & Jude, 1984). Nonlinearity in the measurement model often refers to a nonlinear relationship between binary, ordinal, or nominal item responses and latent factors. In other terms, the conditional distribution of data follows a binomial or multinomial distribution. The nonlinearity in the structural model refers to a nonlinear relationship among latent factors (Kelava & Brandt, 2009). The nonlinear relationship is often estimated by polynomial effects or nonpolynomial effects among latent factors.

Polynomial effects often include quadratic, cubic, and/or interaction effects in educational research (Blozis, 2007). Let $p = 1, \dots, P$ be a person index. Let $\boldsymbol{\omega} = (\theta_{X_p}, \theta_{M_p}, \theta_{Y_p})^T$ be a vector of latent factors. θ_{X_p} is the exogenous latent factor, θ_{M_p} is the endogenous mediator factor, and θ_{Y_p} is the endogenous latent factor. The nonlinear relationship between one exogeneous latent factor θ_{X_p} and one endogenous mediator factor θ_{M_p} is specified as a linear and a quadratic function of θ_{X_p} (McDonald, 1967).

In addition, when θ_M and θ_X are both included in the model as the exogeneous latent factors to predict the endogenous latent factor θ_{Y_p} , it is important to simultaneously estimate the

two exogenous latent factors' quadratic effect as well as their interaction effect to reduce the overestimation of the interaction effect (Harring, Weiss, & Hsu, 2012).

$$\theta_{Y_p} = \beta_1^{\theta_{Y_p}} + \beta_2^{\theta_{Y_p}} \theta_{X_p} + \beta_3^{\theta_{Y_p}} \theta_{X_p}^2 + \beta_4^{\theta_{Y_p}} \theta_{M_p} + \beta_5^{\theta_{Y_p}} \theta_{M_p}^2 + \beta_6^{\theta_{Y_p}} \theta_{X_p} \theta_{M_p} + \delta_p^{\theta_{Y_p}} \quad (1)$$

In the quadratic and interaction model, when θ_{X_p} and θ_{M_p} increase, the expected score of θ_{Y_p} ($E(\theta_{Y_p})$) initially increases but declines after a specific point.

$$E(\theta_{Y_p}) = \beta_1^{\theta_{Y_p}} + \beta_2^{\theta_{Y_p}} \theta_{X_p} + \beta_3^{\theta_{Y_p}} \theta_{X_p}^2 + \beta_4^{\theta_{Y_p}} \theta_{M_p} + \beta_5^{\theta_{Y_p}} \theta_{M_p}^2 + \beta_6^{\theta_{Y_p}} \theta_{X_p} \theta_{M_p} \quad (2)$$

Parameters $\beta_1^{\theta_{Y_p}}$ through $\beta_6^{\theta_{Y_p}}$ govern the direction of the quadratic curve (Sit, Poulin-Costello, & Bergerud, 1994). $\beta_1^{\theta_{Y_p}}$ is the intercept, and $\beta_2^{\theta_{Y_p}}$ and $\beta_4^{\theta_{Y_p}}$ are linear effects of θ_X and θ_M , respectively. $\beta_3^{\theta_{Y_p}}$ and $\beta_5^{\theta_{Y_p}}$ are quadratic effects of θ_{X_p} and θ_{M_p} , respectively. $\beta_6^{\theta_{Y_p}}$ is the interaction effect. $\delta_p^{\theta_{Y_p}}$ is a residual assumed to follow a normal distribution with a mean of 0 and a residual variance of $\sigma_{\delta_{Y_p}}^2$ on the diagonal.

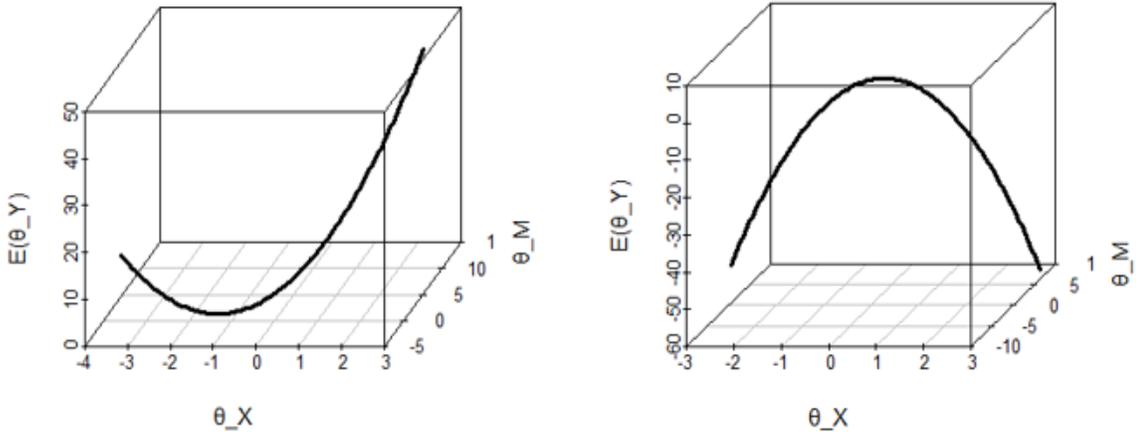


Figure 1: Polynomial nonlinear relationships between exogeneous latent factors (θ_{X_p} , θ_{M_p}) and endogenous latent factor (θ_{Y_p}).

Nonpolynomial nonlinearity includes a broader range of nonlinear functions, such as the exponential function, power function, logarithmic function, cosine function, sine function, and several other nonlinear functions (Sit, Poulin-Costello, & Bergerud, 1994). This study aims to apply the generalized logistic function and sine function in the structural model to explore their nonlinear relationship among latent factors.

The generalized logistic function is one type of exponential function that has been commonly used in item response theory (Hambleton, Swaminathan, & Rogers, 1991) to estimate a nonlinear relationship between binary or polytomous observed responses and latent factors. It posits that when θ_{X_p} increases, θ_{M_p} increases or decreases and remains the same after reaching the asymptote.

$$\theta_{M_p} = \frac{\beta_0^{\theta_{M_p}}}{1 + \exp(\beta_1^{\theta_{M_p}} - \beta_2^{\theta_{M_p}} \theta_{X_p})} + \delta_p^{\theta_{M_p}} \quad (3)$$

When θ_M and θ_X are both included in the model, the θ_{Y_p} is predicted as follows:

$$\theta_{Y_p} = \frac{\beta_3^{\theta_{Y_p}}}{1 + \exp(\beta_4^{\theta_{Y_p}} - \beta_5^{\theta_{Y_p}} \theta_{X_p})} + \frac{\beta_6^{\theta_{Y_p}}}{1 + \exp(\beta_7^{\theta_{Y_p}} - \beta_8^{\theta_{Y_p}} \theta_{M_p})} + \delta_p^{\theta_{Y_p}} \quad (4)$$

$\beta_0^{\theta_{M_p}}$, $\beta_3^{\theta_{Y_p}}$ and $\beta_6^{\theta_{Y_p}}$ are asymptotes, $\beta_1^{\theta_{M_p}}$, $\beta_4^{\theta_{Y_p}}$ and $\beta_7^{\theta_{Y_p}}$ are the amount of change, and

$\beta_2^{\theta_{M_p}}$, $\beta_5^{\theta_{Y_p}}$ and $\beta_8^{\theta_{Y_p}}$ are the rates of change (Sit et al., 1994). The plots in Figure 2 show the curvilinearity between θ_{X_p} , θ_{M_p} , and $E(\theta_Y)$ generated from the generalized logistic function.

$$E(\theta_{Y_p}) = \frac{\beta_3^{\theta_{Y_p}}}{1 + \exp(\beta_4^{\theta_{Y_p}} - \beta_5^{\theta_{Y_p}} \theta_{X_p})} + \frac{\beta_6^{\theta_{Y_p}}}{1 + \exp(\beta_7^{\theta_{Y_p}} - \beta_8^{\theta_{Y_p}} \theta_{M_p})} \quad (5)$$

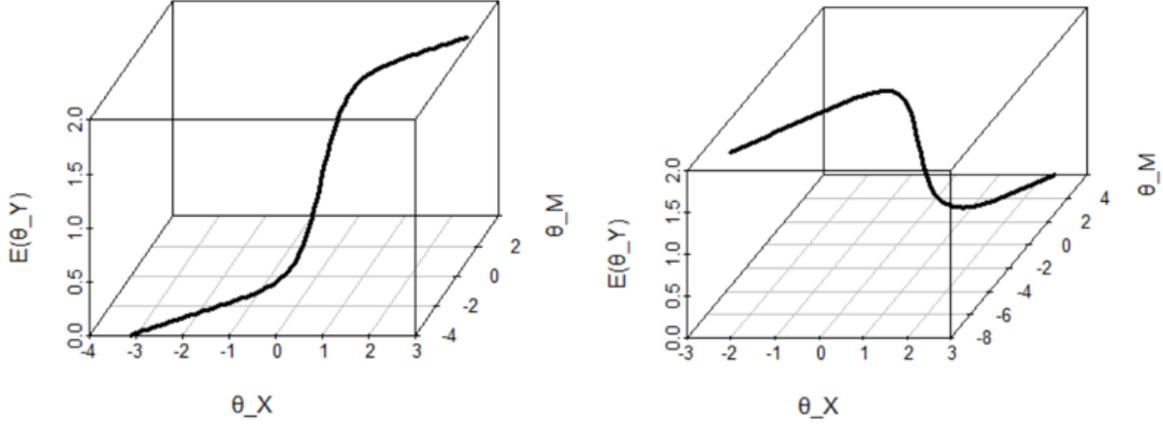


Figure 2: Exponential nonlinear relationships between exogenous latent factors ($\theta_{X_p}, \theta_{M_p}$) and endogenous latent factor (θ_{Y_p}).

The sine function is another nonpolynomial function that is investigated in this study. Sine functions have been applied to estimate periodic curves, such as sound and light waves. These functions help capture the bimodal nonlinear curve between the exogenous latent factor and endogenous mediator factor in the structural model. The sine function is expressed as follows:

$$\theta_{M_p} = \beta_0^{\theta_{M_p}} + \beta_1^{\theta_{M_p}} \sin(\beta_2^{\theta_{M_p}} \theta_{X_p}) + \delta_p^{\theta_{M_p}} \quad (6)$$

The sine curve is expressed as two exogenous latent factors θ_X and θ_M :

$$\theta_{Y_p} = \beta_3^{\theta_{Y_p}} + \beta_4^{\theta_{Y_p}} \sin(\beta_5^{\theta_{Y_p}} \theta_{X_p}) + \beta_6^{\theta_{Y_p}} \sin(\beta_7^{\theta_{Y_p}} \theta_{M_p}) + \delta_p^{\theta_{Y_p}} \quad (7)$$

$\beta_0^{\theta_{M_p}}, \beta_3^{\theta_{Y_p}}$ is the intercept, $\beta_1^{\theta_{M_p}}, \beta_4^{\theta_{Y_p}}$ and $\beta_6^{\theta_{Y_p}}$ govern amount of change, and $\beta_2^{\theta_{M_p}}, \beta_5^{\theta_{Y_p}}$ and $\beta_7^{\theta_{Y_p}}$ govern the rate of change. Figure 3 shows the curvilinearity among $\theta_{X_p}, \theta_{M_p}$, and $E(\theta_{Y_p})$ governed by the sine function.

$$E(\theta_{Y_p}) = \beta_3^{\theta_{Y_p}} + \beta_4^{\theta_{Y_p}} \sin(\beta_5^{\theta_{Y_p}} \theta_{X_p}) + \beta_6^{\theta_{Y_p}} \sin(\beta_7^{\theta_{Y_p}} \theta_{M_p}) \quad (8)$$

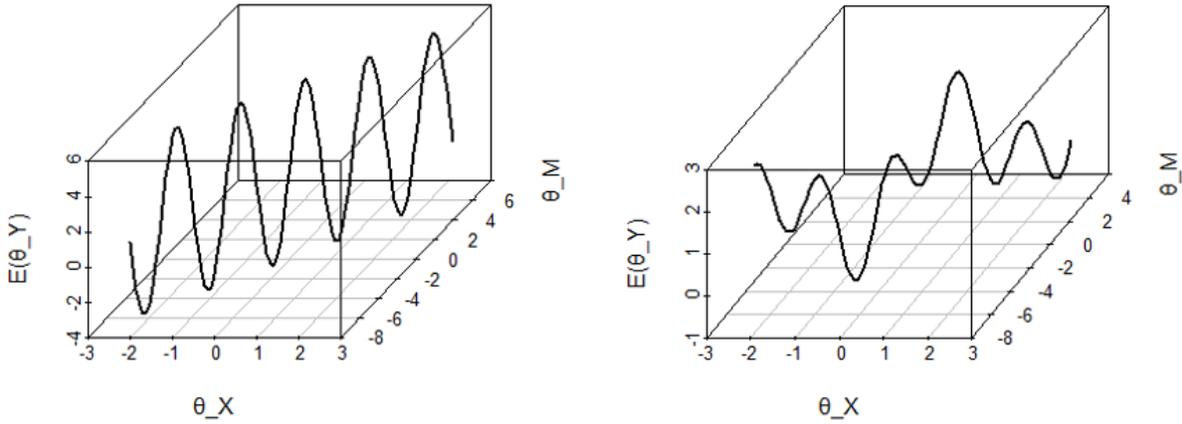


Figure 3: Sinusoidal nonlinear relationships between exogenous latent factors ($\theta_{X_p}, \theta_{M_p}$) and endogenous latent factor (θ_{Y_p}).

Statement of Problems

Most of the nonlinear relations in the latent structural models are estimated within the parametric framework. Two critical limitations of the parametric approach are discussed below.

The first limitation is the normality assumption on latent factors and measurement error (Song, Xia, & Lee, 2009; Song, Pan, Kwok, Vandenput, Ohlsson, & Leung, 2010; Chow, Tang, Yuan, Song, & Zhu, 2011; Yang, Dunson & Baird, 2010). It is common to violate the normality assumption on latent factors in practice. For instance, latent variables in research on rare-event traits, such as a person's tendency to abuse substances, are likely to be nonnormally distributed. Furthermore, when data are nonnormally distributed or heterogeneously distributed, parametric

estimation can lead to unreliable estimation of latent variables and biased parameter estimates (West, Finch & Curran, 1995; Hu, Bentler, & Kano, 1992). Several robust methods were developed in the last decade to relax the normality assumption on latent factors or measurement error, such as the multivariate t -distribution (Lee, 2007), unconstrained approach (Marsh, Wen, & Hau, 2004; Kelave & Brandt, 2009), and quasi-maximum likelihood approach (QML) (Klein & Muthen, 2007). Unfortunately, greater laxity of the normality assumption may reduce the power to detect nonlinear effects (Kelave & Brandt, 2014).

The second limitation is that the parametric method requires prior specification of the functional form of nonlinear relations (e.g., quadratic, cubic) (Bauer, 2005). However, researchers do not have any prior knowledge of the true relations among latent factors in structural models. Therefore, most researchers and practitioners prefer to fit a polynomial function to test the nonlinearity in the structural model due to its ease of computation. The quadratic term is computed by multiplying the exogenous latent factor itself ($\theta_X * \theta_X$), and the interaction term is computed as the product of two exogenous latent factors ($\theta_X * \theta_M$). The problem is that most of the polynomial functions are not sufficient to capture the true nonlinear relations among latent factors.

A small regression simulation is developed to test whether a quadratic effect is statistically significant when the true nonlinear relationship between the independent variable (X) and the dependent variable (Y) is generated from the generalized logistic regression function and the sine function as Figure 4.

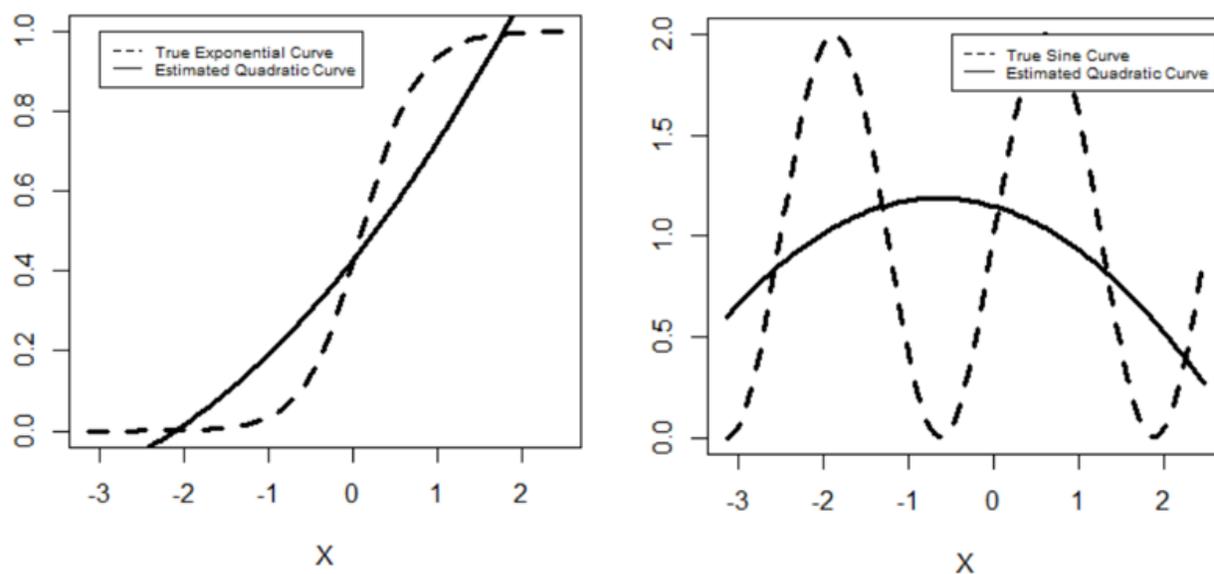


Figure 4: Fitting polynomial nonlinear relations in the nonpolynomial nonlinear relations.

Table 1 lists a statistically significant linear effect and a quadratic effect in the regression analysis when X and Y follow an exponential curve and a bimodal curve.

Table 1: Testing the Quadratic Effect within Nonpolynomial Nonlinearity

	Exponential Data				Sine Data			
	Estimate	SD	p		Estimate	SD	p	
Linear	0.266	0.009	0.000	***	-0.122	0.04	0.007	**
Quadratic	0.030	0.005	0.000	***	-0.094	0.02	0.001	**

The polynomial regression results show that a statistically significant quadratic curve did not represent a true nonlinear curve in most situations. Fitting the polynomial function in nonpolynomial-related data leads to biased parameter estimations and incorrectly inferred results.

Based on these two critical limitations, I propose to apply the semiparametric approach in the latent structural model to relax the normality assumption on both the measurement error and latent factors. More importantly, the semiparametric method does not require a prespecified

functional form of nonlinear relations among latent factors. It provides more flexibility for the selection of nonlinear functions when prior knowledge is not available.

Purpose of the Study

The purpose of this study is to recover the nonlinearity among latent factors using semiparametric Bayesian approach. Specifically, a graded response model (2PL-GRM) is used in the measurement model to account for the nonlinearity between ordered categorical responses and latent factors. The parameters of items (discrimination α , difficulty b) and individuals (latent scores θ) are estimated by the parametric Bayesian approach. A Markov Chain Monte Carlo (MCMC) algorithm is developed for posterior computation. On the other hand, the structural model is estimated by both the parametric and nonparametric Bayesian approach (Ferguson, 1973; Ishwaran & Zarepour, 2000; Ishwaran & James, 2001). Within the nonparametric Bayesian approach, the nonlinear relationships (coefficients β) are estimated via a truncated approximation of the Dirichlet process (DP) prior with a stick-breaking procedure in blocked Gibbs sampling and an MCMC algorithm.

Research Hypotheses

1. The semiparametric Bayesian approach better recovers the polynomial nonlinear curve than the parametric Bayesian approach.
2. The semiparametric Bayesian approach better recovers the exponential nonlinear curve than the parametric Bayesian approach.
3. The semiparametric Bayesian approach better recovers the sinusoidal nonlinear curve than the parametric Bayesian approach.
4. The semiparametric Bayesian approach captures the true exponential nonlinear curve with a polynomial nonlinear function.

5. The semiparametric Bayesian approach captures the true sinusoidal nonlinear curve with a polynomial nonlinear function.

Research Questions

1. What are the differences between the true polynomial nonlinear curve and the estimated nonlinear curve with the parametric Bayesian approach?
2. What are the differences between the true polynomial nonlinear curve and the estimated nonlinear curve with the semiparametric Bayesian approach when latent groups of DP prior are small, medium, and large?
3. What are the differences between the true exponential nonlinear curve and the estimated nonlinear curve with the parametric Bayesian approach?
4. What are the differences between the true exponential nonlinear curve and the estimated nonlinear curve with the semiparametric Bayesian approach when latent groups of DP prior are small, medium, and large?
5. What are the differences between the true sinusoidal nonlinear curve and the estimated nonlinear curve with the parametric Bayesian approach?
6. What are the differences between the true sinusoidal nonlinear curve and the estimated nonlinear curve with the semiparametric Bayesian approach when latent groups of DP prior are small, medium, and large?
7. How well is the true exponential nonlinear curve recovered by the polynomial nonlinear functions with the semiparametric Bayesian approach when latent groups of DP prior are small, medium, and large?

8. How well is the true sinusoidal nonlinear curve recovered by the polynomial nonlinear functions with the semiparametric Bayesian approach when latent groups of DP prior are small, medium, and large?

Plan of the Study

The rest of this study is organized as follows. Chapter Two reviews the parametric and semiparametric approaches applied in earlier studies and the semiparametric structural equation model. Chapter Three develops a simulation study to investigate research hypotheses. The results are presented in Chapter Four, followed by the conclusion and implications of results in Chapter Five.

Chapter 2: Literature Review

The parametric approach assumes that the data come from a population that follows a probability distribution with a fixed set of parameters (Moses, 1952). The nonparametric approach assumes that the data are not generated from a parametric family but rather from an unknown density (Ferguson, 1973). Because the entire density is unknown, the number of parameters is assumed to be infinite. Within the nonparametric framework, Ferguson (1973) introduced the Dirichlet process (DP) as a random probability measure to model the unknown density. Ishwaran and Zarepour (2002) developed the truncated approximation of DP to stimulate the convergence of Markov chains. Ishwaran and James (2001) developed the stick-breaking priors and blocked Gibbs samplers to support the truncated DP prior to sampling the posterior distribution of parameters.

A semiparametric approach combines a nonparametric component involving a portion of the parameters and a parametric component for the other portion of parameters in the model (Ruppert, Wand, & Carroll, 2009). The semiparametric Bayesian method defines the likelihood function of data in the same way as the parametric method has defined it. The difference is in defining the prior distributions for parameters that are assumed to generate from an unknown density. Instead of directly assigning a normal distribution or a conjugate prior distribution (e.g., beta, gamma), the semiparametric method uses a random probability measure, the truncated DP, as the prior to positing in a Gaussian mixture model. Therefore, the parameters' posterior samplings are estimated based on multiple latent groups of priors and a likelihood function. Blocked Gibbs sampling and the MCMC algorithm are used to stimulate the convergence of posterior sampling.

Parametric Approaches in Estimating Nonlinear Relations

In the last two decades, nonlinear relationships among latent factors have been widely assessed by polynomial nonlinear functions within the parametric framework (Kenny & Judd, 1984; Bollen & Paxton, 1998; Jaccard & Wan, 1995; Kelava & Brandt, 2009, 2014; Klein & Moosbrugger, 2000). For example, McDonald (1967) found that a nonlinear model with a quadratic function on an exogeneous factor performs better than a linear model with two exogeneous factors. He extended his research to include the interaction effect among exogeneous latent factors (McDonald, 1967a, 1967b, 1967c). Subsequently, Kenny and Judd (1984) proposed the product indicator approach in a nonlinear structural model to form the latent interaction factor by taking the product of the indicators of two exogenous latent factors and using them as manifest variables for the latent interaction factor (Bollen & Paxton, 1998; Ping, 1995). However, this approach is only valid for normally distributed and centered data (Arminger & Muthen, 1998). When data are skewed, the misspecified variance and covariance in the structural equation leads to biased estimations of parameters. Considering the limitations of the constrained product indicator approach, Klein and Moosbrugger (2000) developed the latent moderated structural equations approach (LMS), in which a maximum likelihood estimator estimates the conditional means and covariances of latent factors through an approximated finite mixture distribution. LMS does not need an interaction term but still assumes normally distributed observed data. Arminger and Muthen (1998) and Lee (2007) proposed the Bayesian method to obtain more accurate inferences without relying on asymptotic assumptions (Zeger & Karim, 1991). Marsh et al., (2004) proposed the unconstrained approach to apply in normal and nonnormally distributed data. However, this approach is only robust to estimating one interaction effect in a small model (Kelave & Brandt, 2009).

In summary, most of the parametric approaches tend to relax the normality assumption on latent factors or measurement error. However, the relaxation of distributional assumptions may lead to biased parameter estimations and a reduction in the power to detect the nonlinear relationship under a parametric framework (Kelava & Brandt, 2014). Therefore, researchers have begun to consider using nonparametric or semiparametric approaches to estimate nonnormally distributed density and/or nonlinear relationships (Robins, Rotnitzky, & Zhao, 1995; Kleinman & Ibrahim, 1998; Song, Xia, Pan, & Lee, 2011; Xia & Gou, 2016; Song et al., 2010; Lee & Song, 2012; Chow, Tang, Yuan, Song, & Zhu, 2011; Yang et al., 2007; Fahrmei & Raah, 2007; Kelava & Brandt, 2014).

Semiparametric Approaches in Estimating Nonnormal Density

Many researchers have applied the truncated approximation of DP with stick-breaking prior, Gibbs sampling, and the MCMC algorithm in a variety of models to address the issues related to nonnormally distributed residuals and latent factors. Kleinman and Ibrahim (1998) applied the semiparametric Bayesian approach in the generalized linear mixed model (GLMM) by nonparametrically modeling both the fixed effect and the random effect. Fahrmei and Raah (2006) applied the Bayesian semiparametric approach for mixed ordinal and continuous responses to nonparametrically modeling the covariate's linear effect and interaction effect. Both Young et al., (2010) and Lee (2007) proposed the semiparametric hierarchical model to nonparametrically model exogeneous latent factors. However, Young et al., (2010) differed in their approach by allowing a mix of categorical and continuous observed data in the measurement model. Song et al. (2010) proposed the semiparametric Bayesian approach to nonparametrically estimate nonnormally distributed residual errors in the measurement equation. Kelava and Brandt (2014) applied a semiparametric Bayesian approach in a multilevel SEM to

handle nonnormally distributed data. Xia and Gou (2014) applied the semiparametric Bayesian method to model the distribution of intercepts and covariance of parameters in structural equations.

Semiparametric Approaches in Estimating Nonlinearity

However, few researchers have applied the semiparametric approach to investigate nonlinearity among latent factors. Bauer (2005) proposed the finite mixture of an SEM with frequentist estimation to test nonlinear relations among latent factors. Song, Lu, Cai, & Hak-Sing Ip (2013) proposed a generalized semiparametric SEM for mixed observed responses to model different functional forms among latent factors. Instead of using DP priors, a Bayesian P-spline approach and MCMC methods are developed to estimate the linear and nonlinear relations. However, the nonlinear function in Song et al. (2013) is limited by the quadratic effect.

This study differs from previous studies in two aspects. First, the semiparametric Bayesian approach is applied in the SEM to explore a variety of nonlinear functional forms, testing not only polynomial functions (e.g., quadratic), as earlier studies have done, but also nonpolynomial functions (e.g., exponential, sine) and how well the semiparametric approach captures the nonlinear relations in each functional form at different truncation levels. The truncation levels refers to the discrete latent groups developed by stick-breaking procedure. Most of the earlier studies stayed at lower truncation levels (e.g., 5, 10). This study explores whether a higher truncation level (e.g., 200) provides a better recovery of nonlinear relations. Second, this study provides a meaningful extension to explore whether the semiparametric Bayesian approach captures nonpolynomial nonlinear relations based on a misspecified polynomial functional form at different truncation levels.

Semiparametric SEM (SSEM)

SEM consists of two components: a measurement model and a structural model. The measurement model of SEM quantifies the relationships between latent and observed variables as well as provides the ability to partition measurement error out of the analysis by the use of item-specific residual variances (Lee, 2007). Relationships among latent variables are modeled by the second component, the structural model. The structural model regresses the exogenous latent variables on the endogenous latent variables.

The SSEM also includes a measurement equation and a structural equation. The major differences between SSEM and classical linear SEM are the distribution of residual variance of data and the distribution of endogenous latent factors. Let i represent an item and c represent a category within an item. In linear SEM, the distribution of y_p is assumed to follow a multivariate normal distribution, and the distribution of latent factors is assumed to follow a multivariate normal distribution with a mean vector $\mathbf{0}$ and a residual covariance matrix Ψ . In contrast, the proposed SSEM model assumes y_p follows a multinomial distribution, and the endogenous latent factors and their relations with the exogenous latent factors are modeled by a nonparametric approach.

Measurement model. In social sciences, observed responses measured by a 3- or 5-point Likert scale are treated as polytomous responses. Therefore, IRT models are appropriate to estimate the nonlinearity in the measurement model under the SEM framework. In this study, the GRM (Samejima, 1997) is used for modeling relationships between ordered response categories and latent factors.

To demonstrate, let X_{pi_X} be the observable variable corresponding to the responses for each individual p and item i_X ($i_X = 1, \dots, I_X$), measuring exogeneous latent factor θ_{X_p} with

categories $c_{i_X} = 1, \dots, C_{i_X}$. Let M_{pi_M} be the observable variable corresponding to the responses for each individual p and item i_M ($i_M = 1, \dots, I_M$), measuring endogenous mediator factor θ_{M_p} with categories $c_{i_M} = 1, \dots, C_{i_M}$. Let Y_{pi_Y} be the observable variable corresponding to the responses for each individual p and item i_Y ($i_Y = 1, \dots, I_Y$), measuring endogenous latent factor θ_{Y_p} with categories $c_{i_Y} = 1, \dots, C_{i_Y}$.

Samejima (1969) proposed a 2PL-GRM to specify the cumulative probability of an individual p of responding to an item i in a given category and above as the following:

$$\begin{aligned} \text{logit}(P(X_{pi_X} \geq c_{i_X}) | \theta_{X_p}, a_{i_X}, b_{i_X c_{i_X}}) &= a_{i_X}(\theta_{X_p} - b_{i_X c_{i_X}}) \\ \text{logit}(P(M_{pi_M} \geq c_{i_M}) | \theta_{M_p}, a_{i_M}, b_{i_M c_{i_M}}) &= a_{i_M}(\theta_{M_p} - b_{i_M c_{i_M}}) \quad (9) \\ \text{logit}(P(Y_{pi_Y} \geq c_{i_Y}) | \theta_{Y_p}, a_{i_Y}, b_{i_Y c_{i_Y}}) &= a_{i_Y}(\theta_{Y_p} - b_{i_Y c_{i_Y}}) \end{aligned}$$

a_{i_X} , a_{i_M} , and a_{i_Y} are the discrimination parameters that describe how well an item can distinguish people with high ability from those with low ability. The higher the discrimination, the more informative the item is. $b_{i_X c_{i_X}}$, $b_{i_M c_{i_M}}$, and $b_{i_Y c_{i_Y}}$ are difficulty thresholds, indicating the location on the ability scale where there is a 50 – 50 chance of responding in category c and higher on an item i .

The endogenous latent factors θ_{M_p} and θ_{Y_p} are assumed from an unknown density and are estimated nonparametrically. For model identification purposes, the exogenous latent factor θ_{X_p} is assumed to follow a normal distribution with a mean of 0 and a variance of 1.

$$\theta_{X_p} \sim N(0, 1) \quad (10)$$

In the measurement equation, the discrimination parameter of the first item measuring θ_{M_p} is fixed as 1, and the first difficulty threshold of the first item measuring θ_{M_p} is fixed as 0.

$$a_{1M} = 1, b_{1M11M} = 0 \quad (11)$$

The same parameter identification is for θ_{Y_p} as well. The discrimination parameter of the first item measuring θ_{Y_p} is fixed as 1, and the first difficulty threshold of the first item measuring θ_{Y_p} is fixed as 0.

$$a_{1Y} = 1, b_{1Y11Y} = 0 \quad (12)$$

The probability of an examinee p responding at a given category c_{iX} conditional on the item parameters and the latent factor θ_{X_p} equals the conditional probability for an individual p on an item i being in category c and higher minus the conditional probability for an individual p on an item i being in a category higher than c .

$$\begin{aligned} P(X_{piX} = c \mid \theta_{X_p}, a_{iX}, b_{iXc_{iX}}) \\ &= P(X_{piX} \geq c \mid \theta_{X_p}, a_{iX}, b_{iXc_{iX}}) \\ &\quad - P(X_{piX} \geq c + 1 \mid \theta_{X_p}, a_{iX}, b_{iX(c_{iX}+1)}) \end{aligned} \quad (13)$$

The probability of an examinee p responding at a given category c_{iM} conditional on the item parameters and the latent factor θ_{M_p} is as follows:

$$\begin{aligned} P(M_{piM} = c \mid \theta_{M_p}, a_{iM}, b_{iMc_{iM}}) \\ &= P(M_{piM} \geq c \mid \theta_{M_p}, a_{iM}, b_{iMc_{iM}}) \\ &\quad - P(M_{piM} \geq c + 1 \mid \theta_{M_p}, a_{iM}, b_{iM(c_{iM}+1)}) \end{aligned} \quad (14)$$

In addition, the probability of an examinee p responding at a given category c_{iY} conditional on the item parameters and the latent factor θ_{Y_p} is as follows:

$$\begin{aligned}
P(Y_{pi_Y} = c \mid \theta_{Y_p}, a_{i_Y}, b_{i_Y c_{i_Y}}) \\
&= P(Y_{pi_Y} \geq c \mid \theta_{Y_p}, a_{i_Y}, b_{i_Y c_{i_Y}}) \\
&\quad - P(Y_{pi_Y} \geq c + 1 \mid \theta_{Y_p}, a_{i_Y}, b_{i_Y(c_{i_Y}+1)})
\end{aligned} \tag{15}$$

Equations (13) - (15) are identified by restricting the probability of a response at or above the lowest category to 1, the probability of a response above the highest category to 0, and threshold $b_{i_X c_{i_X}} \geq b_{i_X(c_{i_X}+1)}$, $b_{i_M c_{i_M}} \geq b_{i_M(c_{i_M}+1)}$, and $b_{i_Y c_{i_Y}} \geq b_{i_Y(c_{i_Y}+1)}$. The conditional probability for an observed response X_{pi_X} is expressed as follows:

$$P(X_{pi_X} \mid \theta_{X_p}, a_{i_X}, b_{i_X c_{i_X}}) = \prod_{c=1}^{c_{i_X}} P(X_{pi_X} = c \mid \theta_{X_p}, a_{i_X}, b_{i_X c_{i_X}})^{F(X_{pi_X}=c)} \tag{16}$$

The conditional probability for an observed response M_{pi_M} is expressed as follows:

$$P(M_{pi_M} \mid \theta_{M_p}, a_{i_M}, b_{i_M c_{i_M}}) = \prod_{c=1}^{c_{i_M}} P(M_{pi_M} = c \mid \theta_{M_p}, a_{i_M}, b_{i_M c_{i_M}})^{F(M_{pi_M}=c)} \tag{17}$$

The conditional probability for an observed response Y_{pi_Y} is expressed as follows:

$$P(Y_{pi_Y} \mid \theta_{Y_p}, a_{i_Y}, b_{i_Y c_{i_Y}}) = \prod_{c=1}^{c_{i_Y}} P(Y_{pi_Y} = c \mid \theta_{Y_p}, a_{i_Y}, b_{i_Y c_{i_Y}})^{F(Y_{pi_Y}=c)} \tag{18}$$

where F is the indicator function that equals 1 when responses for an individual p on an item i is category c or otherwise equals 0.

Structural model. Let $\mathbf{L} = (L_1, L_2, L_3, \dots, L_G)^T$ be a latent classification variable and let $g = 1, \dots, G$ be the groups embedded within \mathbf{L} . Let $\boldsymbol{\beta}^{\theta_{M_p}}$ be a vector of regression coefficients between θ_{X_p} and θ_{M_p} , $\boldsymbol{\beta}^{\theta_{M_p}} = (\beta_0^{\theta_{M_p}}, \beta_1^{\theta_{M_p}}, \beta_2^{\theta_{M_p}})$. Let $\boldsymbol{\beta}^{\theta_{Y_p}}$ be a vector of regression coefficients between θ_{X_p} , θ_{M_p} , and θ_{Y_p} , $\boldsymbol{\beta}^{\theta_{Y_p}} = (\beta_3^{\theta_{Y_p}}, \dots, \beta_8^{\theta_{Y_p}})$. The semiparametric approach

in theory posits a nonfinite mixture of the Gaussian model in the structural model where each parameter varies across the latent groups L_g . Within each latent group g , the conditional distribution of θ_{Y_p} and θ_{M_p} , given parameters, are assumed to follow a normal distribution and are respectively listed as follows:

$$\begin{aligned} (\theta_{M_p} | \boldsymbol{\beta}_g^{\theta_{M_p}}, \theta_{X_p}) &\sim N(u_p^{\theta_{M_p}}, \delta_{pg}^{\theta_{M_p}}) \\ (\theta_{Y_p} | \boldsymbol{\beta}_g^{\theta_{Y_p}}, \theta_{X_p}, \theta_{M_p}) &\sim N(u_p^{\theta_{Y_p}}, \delta_{pg}^{\theta_{Y_p}}) \end{aligned} \quad (19)$$

The residual variance $\delta_{pg}^{\theta_{M_p}}$ is assumed to follow a normal distribution with a mean of 0 and a variance of 2.

$$\delta_{pg}^{\theta_{M_p}} \sim N(0, 2) \quad (20)$$

The residual variance $\delta_{pg}^{\theta_{Y_p}}$ is assumed to follow a normal distribution with a mean of 0 and a variance of 2.

$$\delta_{pg}^{\theta_{Y_p}} \sim N(0, 2) \quad (21)$$

The conditional means of θ_{M_p} and θ_{Y_p} are defined by polynomial, exponential, and sine functional forms within each latent group as follows:

$$\begin{aligned} u_p^{\theta_{M_p}} &= \beta_{0g}^{\theta_{M_p}} + \beta_{1g}^{\theta_{M_p}} \theta_{X_p} + \beta_{2g}^{\theta_{M_p}} \theta_{X_p}^2 \\ u_p^{\theta_{Y_p}} &= \beta_{3g}^{\theta_{Y_p}} + \beta_{4g}^{\theta_{Y_p}} \theta_{X_p} + \beta_{5g}^{\theta_{Y_p}} \theta_{X_p}^2 + \beta_{6g}^{\theta_{Y_p}} \theta_{M_p} + \beta_{7g}^{\theta_{Y_p}} \theta_{M_p}^2 \\ &\quad + \beta_{8g}^{\theta_{Y_p}} \theta_{X_p} \theta_{M_p} \end{aligned} \quad (22)$$

$$u_p^{\theta_{M_p}} = \frac{\beta_{0g}^{\theta_{M_p}}}{1 + \exp(\beta_{1g}^{\theta_{M_p}} - \beta_{2g}^{\theta_{M_p}} \theta_{X_p})} \quad (23)$$

$$u_p^{\theta_{Yp}} = \frac{\beta_{3g}^{\theta_{Yp}}}{1 + \exp(\beta_{4g}^{\theta_{Yp}} - \beta_{5g}^{\theta_{Yp}} \theta_{Xp})} + \frac{\beta_{6g}^{\theta_{Yp}}}{1 + \exp(\beta_{7g}^{\theta_{Yp}} - \beta_{8g}^{\theta_{Yp}} \theta_{Mp})}$$

$$u_p^{\theta_{Mp}} = \beta_{0g}^{\theta_{Mp}} + \beta_{1g}^{\theta_{Mp}} \sin(\beta_{2g}^{\theta_{Mp}} \theta_{Xp})$$

$$u_p^{\theta_{Yp}} = \beta_{3g}^{\theta_{Yp}} + \beta_{4g}^{\theta_{Yp}} \sin(\beta_{5g}^{\theta_{Yp}} \theta_{Xp}) + \beta_{6g}^{\theta_{Yp}} \sin(\beta_{7g}^{\theta_{Yp}} \theta_{Mp})$$
(24)

Based on a suggestion of Ishwaran and James (2001), a truncation approximation of DP with stick-breaking prior is chosen to define the mean of θ_{Mp} and θ_{Yp} as follows:

$$F_G \# (u_p^{\theta_{Mp}}) = \sum_{g=1}^G \pi_g S(u_p^{\theta_{Mp}}), \quad 1 \leq G \leq \infty$$
(25)

$$F_G \# (u_p^{\theta_{Yp}}) = \sum_{g=1}^G \pi_g S(u_p^{\theta_{Yp}}), \quad 1 \leq G \leq \infty$$
(26)

$\boldsymbol{\pi} = (\pi_1, \dots, \pi_G)$ is a vector of discrete weighted variables. The stick-breaking procedure is used to define the random weighted variable π_g in the following steps. First, an infinite sequence $\boldsymbol{v} = (v_1, v_2, \dots, v_g)$ is drawn from an independently and identically (*i. i. d.*) distributed beta distribution:

$$v_g \sim^{iid} \text{Beta}(0.5, 0.5), \quad g = 1, 2, \dots, G - 1,$$
(27)

Then, $\pi_1 = v_1$, $\pi_2 = v_2(1 - v_1)$, and so on becomes as follows:

$$\pi_g = v_g \prod_{j=1}^{g-1} (1 - v_j), \quad g = 2, 3, \dots, G,$$
(28)

The stick-breaking prior is defined by ensuring that the sum of the weighted variables across all latent groups is 1 (Sethuraman, 1994):

$$\sum_{g=1}^G \pi_g = 1 \quad (29)$$

Finally, by applying the discrete weighted variable π_g on a continuous real line $S(u_p^{\theta_{M_p}})$ and $S(u_p^{\theta_{Y_p}})$, consecutively, a continuous distribution is transformed as a discrete distribution between 0 and 1. By summing all the density functions together across all latent groups, a cumulative truncated DP of means is defined (Ishwaran & James, 2001).

Semiparametric Bayesian Estimation

The semiparametric Bayesian method follows the Bayesian theorem that was initially proposed by Thomas Bayes (1763), defining a posterior distribution $P(\theta|x)$ by synthesizing the prior distribution $P(\theta)$ and the likelihood of data $P(x|\theta)$ as follows (Levy & Mislvey, 2016):

$$P(\theta|x) = \frac{P(x|\theta) P(\theta)}{\sum P(x|\theta^*) P(\theta^*)} \propto P(x|\theta)P(\theta) \quad (30)$$

The likelihood function is defined as a conditional probability given a latent factor $P(x|\theta)$ (Levy & Mislvey, 2016). A prior distribution is a priori assumption or knowledge made by the researcher that most closely represents the substantive content of the unknown density (Levy & Mislvey, 2016). The posterior distribution $P(\theta|x)$ of a parameter aims to describe the entire density of a parameter, instead of finding a set of values to maximize the likelihood function, as in maximum likelihood (ML) estimation (Lord, 2012). Within the Bayesian approach, parameter estimation and goodness-of-fit statistics are straightforward and easily implemented (Levy & Mislvey, 2016).

The semiparametric Bayesian method defines the likelihood function of data in the same way in which the parametric method has defined it. The difference between the semiparametric Bayesian method and the parametric method is in defining the prior distributions. Instead of

directly assigning a normal distribution or a conjugate prior distribution (e.g., beta, gamma) as a prior on a parameter, the semiparametric method uses a random probability measure, the truncated DP, as the prior to posit in a Gaussian mixture model. Therefore, the parameters' posterior samplings are simulated based on multiple latent groups of priors and likelihood functions. The marginalized posterior estimation is obtained by marginalizing across all the latent groups. Gibbs sampling and the MCMC algorithm are exploited to simulate posterior samplings.

Chapter 3: Method

Data Generation

A simulation study was conducted to investigate the performance of the Bayesian parametric and semiparametric SEM models. Data were simulated with a 3 * 4 experimental design in which the observed responses were generated from the 2PL-GRM model. A total of 12 conditions were simulated, as shown in Table 2.

Table 2: A 3 x 4 Simulation Design

Polynomial Function	Generalized Logistic Function	Sine Function
Parametric Bayesian Semiparametric Approach, Latent Group $G = 5$	Parametric Bayesian Semiparametric Approach, Latent Group $G = 5$	Parametric Bayesian Semiparametric Approach, Latent Group $G = 5$
Semiparametric Approach, Latent Group $G = 20$	Semiparametric Approach, Latent Group $G = 20$	Semiparametric Approach, Latent Group $G = 20$
Semiparametric Approach, Latent Group $G = 200$	Semiparametric Approach, Latent Group $G = 200$	Semiparametric Approach, Latent Group $G = 200$

Each condition was analyzed by the parametric and semiparametric approach. The nonlinear relationship was estimated based on the posterior samplings of coefficient parameters. The nonconvergence rate in each condition was reported. The range of differences between the true nonlinear curves and the estimated nonlinear curves were compared and summarized in the results.

Generating the measurement model. The R (R Core Team, 2014) and RStudio programming environment (RStudio Team, 2015) was used to create simulated items, factors, and observed responses. The sample size was fixed at $N = 100$ within each condition. Three latent factors ($\theta_{X_P}, \theta_{M_P}, \theta_{Y_P}$), 30 items ($i_X = 1, \dots, 10_X, i_M = 1, \dots, 10_M, i_Y = 1, \dots, 10_Y$), and 5 ordered categories ($C = 5$) within each item ($c_{i_X} = 1, \dots, 5_{i_X}, c_{i_M} = 1, \dots, 5_{i_M}, c_{i_Y} = 1, \dots, 5_{i_Y}$) were simulated based on the 2PL-GRM model (Equation (9)) in the measurement model. To

generate $i_X = 1, \dots, 10_X$, the discrimination parameter was generated from $\alpha_{i_X} \sim \text{unif}(0, 0.5)$, the first difficulty threshold followed a normal distribution $b_{i_X 1_{i_X}} \sim N(0, 0.5)$, and the latter difficulty thresholds were simulated as $b_{i_X 2_{i_X}}, \dots, b_{i_X (c-1)_{i_X}} = b_{i_X (c-1)_{i_X}} + \text{unif}(0, 0.5)$. Next, given that the latent mediator factor was assumed to generate from an unknown density, the discrimination parameter of the first item measuring θ_{M_p} was fixed as 1, $\alpha_{1_M} = 1$, for model identification. The discrimination parameter from item 2, $i_M = 2, \dots, 10_M$, followed a normal distribution $\alpha_{2_M}, \dots, \alpha_{10_M} \sim \text{unif}(0.5, 3)$. The first difficulty threshold of the first item measuring θ_{M_p} was fixed as 0, $b_{1_M 1_{1_M}} = 0$, for the model identification. The second difficulty threshold of the first item measuring θ_{M_p} followed a normal distribution $b_{1_M 2_{1_M}} \sim N(0, 0.5)$, and the latter difficulty thresholds of the first item measuring θ_{M_p} were simulated as $b_{1_M 3_{1_M}}, \dots, b_{1_M (c-1)_{1_M}} = b_{1_M (c-1)_{1_M}} + \text{unif}(0, 0.5)$. For items $i_M = 2, \dots, 10_M$, the first difficulty threshold followed a normal distribution $b_{2_M 1_{2_M}}, \dots, b_{10_M 1_{10_M}} \sim N(0, 0.5)$, and the latter difficulty thresholds were simulated as $b_{2_M 2_{2_M}}, \dots, b_{10_M (c-1)_{10_M}} = b_{i_M (c-1)_{i_M}} + \text{unif}(0, 0.5)$. Finally, θ_{Y_p} was assumed to generate from an unknown density as well; therefore, item parameters were fixed in the same way. The discrimination parameter of the first item measuring θ_{Y_p} was fixed as 1, $\alpha_{1_Y} = 1$. Each of the items from $i_Y = 2, \dots, 10_Y$ followed a normal distribution $\alpha_{2_Y}, \dots, \alpha_{10_Y} \sim \text{unif}(0.5, 3)$. The first difficulty threshold of the first item measuring θ_{Y_p} was fixed as 0, $b_{1_Y 1_{1_Y}} = 0$. The second difficulty threshold of the first item followed a normal distribution $b_{1_Y 2_{1_Y}} \sim N(0, 0.5)$, and the latter difficulty thresholds of the first item were simulated as $b_{1_Y 3_{1_Y}}, \dots, b_{1_Y (c-1)_{1_Y}} = b_{1_Y (c-1)_{1_Y}} + \text{unif}(0, 0.5)$. For items $i_Y = 2, \dots, 10_Y$, the first difficulty threshold followed a normal distribution

$b_{2Y1_2Y}, \dots, b_{10Y1_{10Y}} \sim N(0, 0.5)$, and the latter difficulty thresholds were simulated as

$$b_{2Y2_2Y}, \dots, b_{10Y(C-1)_{10Y}} = b_{i_Y(c-1)_{i_Y}} + \text{unif}(0, 0.5).$$

Generating the structural model. Three nonlinear functions (Equations (22)(23), and (24)) were used to generate the expected means of three nonlinear relations among the three latent factors θ_{X_P} , θ_{M_P} , and θ_{Y_P} . Linear, quadratic, and interaction effects were included in the polynomial function to generate the expected score of θ_{M_P} and θ_{Y_P} . An orthogonal polynomial was used on linear and quadratic effects on the exogenous latent factor and the mediator latent factor in Equation (22). The generalized logistic regression and the sine functions were simulated based on Equations (23) and (24). $\beta^{\theta_{M_P}}$ and $\beta^{\theta_{Y_P}}$ were simulated from a normal distribution with a mean of 0 and a variance of 0.5 in each nonlinear function, $\beta^{\theta_{M_P}} \sim N(0, 0.5)$ and $\beta^{\theta_{Y_P}} \sim N(0, 0.5)$. The variance of θ_{M_P} and θ_{Y_P} followed a normal distribution with a mean of 0 and a variance of 2 in each nonlinear function, $\delta_p^{\theta_{M_P}}$ and $\delta_p^{\theta_{Y_P}}$ were fixed as 2.

Data Analysis

An MCMC algorithm was adopted to estimate model parameters, which was implemented in the JAGS software (Plummer, 2003) by using the R2jags package (Su & Yajima, 2012) in the programming environment RStudio (RStudio Team, 2015). The parametric Bayesian method was first performed as the reference approach to estimate the regression coefficients in three nonlinear functions. Next, the truncated DP with the stick-breaking prior and the MCMC algorithm were used to simulate posterior samplings of parameters. There was no difference in the calibration of the discrimination parameter and the difficulty parameters in the measurement model between the parametric and the semiparametric Bayesian approaches in the

SEM. The difference between the two approaches was in defining the prior of the nonlinear coefficients in the structural model.

Study 1. The objective of study 1 was to investigate how well the semiparametric approach captures the true nonlinear relationships among latent factors under the polynomial, exponential, and sine functions.

Parametric Bayesian approach. Let $\boldsymbol{\tau} =$

$(\boldsymbol{\beta}^{\theta_{M_p}}, \boldsymbol{\beta}^{\theta_{Y_p}}, \alpha_{i_X}, \alpha_{i_M}, \alpha_{i_Y}, b_{i_X c_{i_X}}, b_{i_M c_{i_M}}, b_{i_Y c_{i_Y}}, \delta_p^{\theta_{M_p}}, \delta_p^{\theta_{Y_p}})$ be a vector to include all the parameters. The conditional likelihood of the data was as follows:

$$\begin{aligned}
 P(\mathbf{X} | \theta_{X_p}, \theta_{M_p}, \theta_{Y_p}, \boldsymbol{\tau}) &= \prod_{n=1}^N \prod_{i=1}^I P(X_{pi_X} | \theta_{X_p}, \theta_{M_p}, \theta_{Y_p}, \boldsymbol{\tau}) \\
 P(\mathbf{M} | \theta_{X_p}, \theta_{M_p}, \theta_{Y_p}, \boldsymbol{\tau}) &= \prod_{n=1}^N \prod_{i=1}^I P(M_{pi_M} | \theta_{X_p}, \theta_{M_p}, \theta_{Y_p}, \boldsymbol{\tau}) \\
 P(\mathbf{Y} | \theta_{X_p}, \theta_{M_p}, \theta_{Y_p}, \boldsymbol{\tau}) &= \prod_{n=1}^N \prod_{i=1}^I P(Y_{pi_Y} | \theta_{X_p}, \theta_{M_p}, \theta_{Y_p}, \boldsymbol{\tau})
 \end{aligned} \tag{31}$$

Where the observed responses X_{pi_X} , M_{pi_X} , and Y_{pi_X} followed a multinomial distribution:

$$\begin{aligned}
 X_{pi_X} | \theta_{X_p}, \theta_{M_p}, \theta_{Y_p}, \boldsymbol{\tau} &\sim \text{Categorical } P(X_{pi_X} | \theta_{X_p}, \theta_{M_p}, \theta_{Y_p}, \boldsymbol{\tau}) \\
 M_{pi_M} | \theta_{X_p}, \theta_{M_p}, \theta_{Y_p}, \boldsymbol{\tau} &\sim \text{Categorical } P(M_{pi_M} | \theta_{X_p}, \theta_{M_p}, \theta_{Y_p}, \boldsymbol{\tau}) \\
 Y_{pi_Y} | \theta_{X_p}, \theta_{M_p}, \theta_{Y_p}, \boldsymbol{\tau} &\sim \text{Categorical } P(Y_{pi_Y} | \theta_{X_p}, \theta_{M_p}, \theta_{Y_p}, \boldsymbol{\tau})
 \end{aligned} \tag{32}$$

Priors distribution. A prior distribution was assigned to each parameter. For discrimination parameters α_{i_X} , α_{i_M} , and α_{i_Y} , a normal prior distribution with a mean of 1 and a variance of 2 was used in the model. For the first difficulty threshold parameter, $b_{i_X 1_{i_X}}$, $b_{i_M 1_{i_M}}$, and $b_{i_Y 1_{i_Y}}$ followed a normal distribution with a mean of 0 and a variance of 0.5. To impose the

restriction that $b_{i(c-1)} \geq b_{ic}$, the prior distributions of latter difficulty thresholds were specified as follows:

$$\begin{aligned} b_{i_X 2 i_X}, \dots, b_{i_X (c-1) i_X} &= b_{i_X (c-1) i_X} + N(0.1, 0.2)T(0) \\ b_{i_M 2 i_M}, \dots, b_{i_M (c-1) i_M} &= b_{i_M (c-1) i_M} + N(0.1, 0.2)T(0) \\ b_{i_Y 2 i_Y}, \dots, b_{i_Y (c-1) i_Y} &= b_{i_Y (c-1) i_Y} + N(0.1, 0.2)T(0) \end{aligned} \quad (33)$$

The priors of the $\boldsymbol{\beta}^{\theta_{MP}}$ and $\boldsymbol{\beta}^{\theta_{YP}}$ followed a normal distribution with a mean of 0.5 and a variance of 0.1.

$$\begin{aligned} \boldsymbol{\beta}^{\theta_{MP}} &\sim N(0.5, 0.1) \\ \boldsymbol{\beta}^{\theta_{YP}} &\sim N(0.5, 0.1) \end{aligned} \quad (34)$$

The prior for the variance of endogenous latent factor's variance followed a gamma distribution.

$$\begin{aligned} \delta_p^{\theta_{MP}} &\sim \text{gamma}(11, 1) \\ \delta_p^{\theta_{YP}} &\sim \text{gamma}(11, 1) \end{aligned} \quad (35)$$

Posterior distribution. Gibbs sampling and the MCMC algorithm were used to estimate the posterior distribution of each parameter. The posterior distribution for all parameters was defined as follows:

$$\begin{aligned} &P(\theta_{X_P}, \theta_{M_P}, \theta_{Y_P}, \boldsymbol{\tau} | \mathbf{X}, \mathbf{M}, \mathbf{Y}) \\ &= \prod_{n=1}^N \prod_{i=1}^I P(X_{pi_X}, M_{pi_M}, Y_{pi_Y} | \theta_{X_P}, \theta_{M_P}, \theta_{Y_P}, \boldsymbol{\tau}) P(\theta_{X_P}) P(\theta_{M_P}) P(\theta_{Y_P}) \\ &P(\boldsymbol{\beta}^{\theta_{MP}}) P(\boldsymbol{\beta}^{\theta_{YP}}) P(\alpha_{i_X}) P(\alpha_{i_M}) P(\alpha_{i_Y}) P(\delta_p^{\theta_{MP}}) P(\delta_p^{\theta_{YP}}) \\ &\prod_{c=2}^{c_i} P(b_{i_X c_{i_X}}) \prod_{c=2}^{c_i} P(b_{i_M c_{i_M}}) \prod_{c=2}^{c_i} P(b_{i_Y c_{i_Y}}) \end{aligned} \quad (36)$$

Semiparametric Bayesian approach. In the semiparametric Bayesian approach, the truncated DP with the stick-breaking prior was added to the model to define the prior distribution of parameters that were estimated nonparametrically. Let $\boldsymbol{\tau} =$

$(\boldsymbol{\beta}^{\theta_{MP}}, \boldsymbol{\beta}^{\theta_{YP}}, \alpha_{i_X}, \alpha_{i_M}, \alpha_{i_Y}, b_{i_X c_{i_X}}, b_{i_M c_{i_M}}, b_{i_Y c_{i_Y}}, \delta_{pg}^{\theta_{MP}}, \delta_{pg}^{\theta_{YP}}, \boldsymbol{\pi}, \boldsymbol{v})$ be a vector to include all the parameters. The conditional likelihood of the data was the same as the parametric Bayesian approach in Equations (31) and (32). The difference was concentrated on defining the prior of the regression coefficients $\boldsymbol{\beta}^{\theta_{MP}}$ and $\boldsymbol{\beta}^{\theta_{YP}}$ and the variance of the variance of the endogenous latent factors $\sigma_p^{2\theta_{MP}}$ and $\sigma_p^{2\theta_{YP}}$.

Prior distribution. In theory, a truncated DP with the stick-breaking prior develops a Gaussian mixture model to define the prior distribution of $\boldsymbol{\beta}^{\theta_{MP}}, \boldsymbol{\beta}^{\theta_{YP}}, \sigma_p^{2\theta_{MP}}$, and $\sigma_p^{2\theta_{YP}}$.

Within each latent class, the prior of the $\boldsymbol{\beta}_g^{\theta_{MP}}$ and $\boldsymbol{\beta}_g^{\theta_{YP}}$ followed a normal distribution with a mean of 0.5 and a variance of 0.1.

$$\begin{aligned}\boldsymbol{\beta}_g^{\theta_{MP}} &\sim N(0.5, 0.1) \\ \boldsymbol{\beta}_g^{\theta_{YP}} &\sim N(0.5, 0.1)\end{aligned}\tag{37}$$

The prior of the residual variance of endogenous latent factor's variance followed a gamma distribution.

$$\begin{aligned}\delta_{pg}^{\theta_{MP}} &\sim \text{gamma}(11, 1) \\ \delta_{pg}^{\theta_{YP}} &\sim \text{gamma}(11, 1)\end{aligned}\tag{38}$$

The prior of parameter v_g in the stick-breaking procedure was defined by a beta distribution.

$$v_g \sim \text{Beta}(0.5, 0.5)\tag{39}$$

Posterior distribution. Gibbs sampling and the MCMC algorithm were used to estimate the posterior distribution of each parameter. The posterior distribution for all parameters was defined as follows:

$$\begin{aligned}
& P\left(\theta_{X_p}, \theta_{M_p}, \theta_{Y_p}, \boldsymbol{\tau} \mid \mathbf{X}, \mathbf{M}, \mathbf{Y}\right) \\
&= \prod_{n=1}^N \prod_{i=1}^I P\left(X_{pi_X}, M_{pi_M}, Y_{pi_Y} \mid \theta_{X_p}, \theta_{M_p}, \theta_{Y_p}, \boldsymbol{\tau}\right) P\left(\theta_{X_p}\right) P\left(\theta_{M_p}\right) P\left(\theta_{Y_p}\right) \\
& P\left(\boldsymbol{\beta}_g^{\theta_{M_p}}\right) P\left(\boldsymbol{\beta}_g^{\theta_{Y_p}}\right) P\left(\alpha_{i_X}\right) P\left(\alpha_{i_M}\right) P\left(\alpha_{i_Y}\right) P\left(\delta_{pg}^{\theta_{M_p}}\right) P\left(\delta_{pg}^{\theta_{Y_p}}\right) \\
& P\left(\pi_g\right) P\left(v_g\right) \prod_{c=2}^{c_i} P\left(b_{i_X c_{i_X}}\right) \prod_{c=2}^{c_i} P\left(b_{i_M c_{i_M}}\right) \prod_{c=2}^{c_i} P\left(b_{i_Y c_{i_Y}}\right)
\end{aligned} \tag{40}$$

Outcomes. The true polynomial nonlinear curve was calculated as follows:

$$\begin{aligned}
y_{true_{poly}}^{\theta_{M_p}} &= \beta_0^{\theta_{M_p}} + \beta_1^{\theta_{M_p}} \theta_{X_p} + \beta_2^{\theta_{M_p}} \theta_{X_p}^2 \\
y_{true_{poly}}^{\theta_{Y_p}} &= \beta_3^{\theta_{Y_p}} + \beta_4^{\theta_{Y_p}} \theta_{X_p} + \beta_5^{\theta_{Y_p}} \theta_{X_p}^2 + \beta_6^{\theta_{Y_p}} \theta_{M_p} + \beta_7^{\theta_{Y_p}} \theta_{M_p}^2 \\
&+ \beta_8^{\theta_{Y_p}} \theta_{X_p} \theta_{M_p}
\end{aligned} \tag{41}$$

The estimated polynomial nonlinear curve with the semiparametric Bayesian approach was calculated as follows:

$$\begin{aligned}
y_{esti_{poly}}^{\theta_{M_p}} &= \frac{\sum_1^{n.iter} \sum_{g=1}^G \pi_g \left(\beta_{0_g}^{\theta_{M_p}} + \beta_{1_g}^{\theta_{M_p}} \theta_{X_p} + \beta_{2_g}^{\theta_{M_p}} \theta_{X_p}^2\right)}{\# \text{ of iterations}} \\
y_{esti_{poly}}^{\theta_{Y_p}} &= \frac{\sum_1^{n.iter} \sum_{g=1}^G \pi_g \left(\beta_{3_g}^{\theta_{Y_p}} + \beta_{4_g}^{\theta_{Y_p}} \theta_{X_p} + \beta_{5_g}^{\theta_{Y_p}} \theta_{X_p}^2 + \beta_{6_g}^{\theta_{Y_p}} \theta_{M_p} + \beta_{7_g}^{\theta_{Y_p}} \theta_{M_p}^2 + \beta_{8_g}^{\theta_{Y_p}} \theta_{X_p} \theta_{M_p}\right)}{\# \text{ of iterations}}
\end{aligned} \tag{42}$$

The difference was summarized as follows:

$$Diff_{poly}^{\theta_{M_p}} = y_{true_{poly}}^{\theta_{M_p}} - y_{esti_{poly}}^{\theta_{M_p}} \quad (43)$$

$$Diff_{poly}^{\theta_{Y_p}} = y_{true_{poly}}^{\theta_{Y_p}} - y_{esti_{poly}}^{\theta_{Y_p}}$$

The true exponential nonlinear curve was calculated as follows:

$$y_{true_{expo}}^{\theta_{M_p}} = \frac{\beta_0^{\theta_{M_p}}}{1 + \exp(\beta_1^{\theta_{M_p}} - \beta_2^{\theta_{M_p}} \theta_{X_p})}$$

$$y_{true_{expo}}^{\theta_{Y_p}} = \frac{\beta_3^{\theta_{Y_p}}}{1 + \exp(\beta_4^{\theta_{Y_p}} - \beta_5^{\theta_{Y_p}} \theta_{X_p})} + \frac{\beta_6^{\theta_{Y_p}}}{1 + \exp(\beta_7^{\theta_{Y_p}} - \beta_8^{\theta_{Y_p}} \theta_{M_p})} \quad (44)$$

The estimated exponential nonlinear curve with the semiparametric Bayesian approach was calculated as follows:

$$y_{esti_{expo}}^{\theta_{M_p}} = \frac{\sum_1^{n.iter} \sum_{g=1}^G \pi_g \left(\frac{\beta_{0g}^{\theta_{M_p}}}{1 + \exp(\beta_{1g}^{\theta_{M_p}} - \beta_{2g}^{\theta_{M_p}} \theta_{X_p})} \right)}{\# \text{ of iterations}} \quad (45)$$

$$y_{esti_{expo}}^{\theta_{Y_p}} = \frac{\sum_1^{n.iter} \sum_{g=1}^G \pi_g \left(\frac{\beta_{3g}^{\theta_{Y_p}}}{1 + \exp(\beta_{4g}^{\theta_{Y_p}} - \beta_{5g}^{\theta_{Y_p}} \theta_{X_p})} + \frac{\beta_{6g}^{\theta_{Y_p}}}{1 + \exp(\beta_{7g}^{\theta_{Y_p}} - \beta_{8g}^{\theta_{Y_p}} \theta_{M_p})} \right)}{\# \text{ of iterations}}$$

The difference was summarized as follows:

$$Diff_{expo}^{\theta_{M_p}} = y_{true_{expo}}^{\theta_{M_p}} - y_{esti_{expo}}^{\theta_{M_p}} \quad (46)$$

$$Diff_{expo}^{\theta_{Y_p}} = y_{true_{expo}}^{\theta_{Y_p}} - y_{esti_{expo}}^{\theta_{Y_p}}$$

The true sine nonlinear curve was calculated as follows:

$$\begin{aligned}
y_{true_{sine}}^{\theta_{Mp}} &= \beta_0^{\theta_{Mp}} + \beta_1^{\theta_{Mp}} \sin(\beta_2^{\theta_{Mp}} \theta_{Xp}) \\
y_{true_{sine}}^{\theta_{Yp}} &= \beta_3^{\theta_{Yp}} + \beta_4^{\theta_{Yp}} \sin(\beta_5^{\theta_{Yp}} \theta_{Xp}) + \beta_6^{\theta_{Yp}} \sin(\beta_7^{\theta_{Yp}} \theta_{Mp})
\end{aligned} \tag{47}$$

The estimated sine nonlinear curve with the semiparametric Bayesian approach was calculated as follows:

$$\begin{aligned}
y_{esti_{sine}}^{\theta_{Mp}} &= \frac{\sum_1^{n.iter} \sum_{g=1}^G \pi_g (\beta_{0g}^{\theta_{Mp}} + \beta_{1g}^{\theta_{Mp}} \sin(\beta_{2g}^{\theta_{Mp}} \theta_{Xp}))}{\# of iterations} \\
y_{esti_{sine}}^{\theta_{Yp}} &= \frac{\sum_1^{n.iter} \sum_{g=1}^G \pi_g (\beta_{3g}^{\theta_{Yp}} + \beta_{4g}^{\theta_{Yp}} \sin(\beta_{5g}^{\theta_{Yp}} \theta_{Xp}) + \beta_{6g}^{\theta_{Yp}} \sin(\beta_{7g}^{\theta_{Yp}} \theta_{Mp}))}{\# of iterations}
\end{aligned} \tag{48}$$

The difference was summarized as follows:

$$\begin{aligned}
Diff_{sine}^{\theta_{Mp}} &= y_{true_{sine}}^{\theta_{Mp}} - y_{esti_{sine}}^{\theta_{Mp}} \\
Diff_{sine}^{\theta_{Yp}} &= y_{true_{sine}}^{\theta_{Yp}} - y_{esti_{sine}}^{\theta_{Yp}}
\end{aligned} \tag{49}$$

In study 1, the polynomial nonlinear function, the exponential nonlinear function, and the sine nonlinear function were specified in the simulated polynomial, exponential, and sine datasets, respectively. A total of 4 chains, 18,000 iterations per chain, 8000 burn-in, and 100 thin were implemented in the analysis of the polynomial and exponential function. A total of 400 iterations were utilized in the posterior analysis to plot the estimated polynomial curve and the estimated exponential curve. A total of 4 chains, 20,000 iterations per chain, 10,000 burn-in, and 100 thin were implemented in the analysis of the sine function. A total of 400 iterations were used in the posterior analysis to plot the estimated sine curve.

Study 2. The objective of the study 2 was to explore whether the semiparametric approach helps capture the true nonpolynomial nonlinear relations when the polynomial quadratic and interaction effects were prespecified and estimated in the model.

Outcomes. The semiparametric Bayesian approach was used to analyze the posterior samplings of parameters. The true exponential nonlinear curve was calculated as follows:

$$y_{true_{expo}}^{\theta_{M_p}} = \frac{\beta_0^{\theta_{M_p}}}{1 + \exp(\beta_1^{\theta_{M_p}} - \beta_2^{\theta_{M_p}} \theta_{X_p})}$$

$$y_{true_{expo}}^{\theta_{Y_p}} = \frac{\beta_3^{\theta_{Y_p}}}{1 + \exp(\beta_4^{\theta_{Y_p}} - \beta_5^{\theta_{Y_p}} \theta_{X_p})} + \frac{\beta_6^{\theta_{Y_p}}}{1 + \exp(\beta_7^{\theta_{Y_p}} - \beta_8^{\theta_{Y_p}} \theta_{M_p})} \quad (50)$$

The estimated nonlinear curve with a semiparametric Bayesian approach based on a polynomial nonlinear function was calculated as follows:

$$y_{esti_{poly}}^{\theta_{M_p}} = \frac{\sum_1^{n.iter} \sum_{g=1}^G \pi_g (\beta_{0g}^{\theta_{M_p}} + \beta_{1g}^{\theta_{M_p}} \theta_{X_p} + \beta_{2g}^{\theta_{M_p}} \theta_{X_p}^2)}{\# \text{ of iterations}}$$

$$y_{esti_{poly}}^{\theta_{Y_p}} = \frac{\sum_1^{n.iter} \sum_{g=1}^G \pi_g (\beta_{3g}^{\theta_{Y_p}} + \beta_{4g}^{\theta_{Y_p}} \theta_{X_p} + \beta_{5g}^{\theta_{Y_p}} \theta_{X_p}^2 + \beta_{6g}^{\theta_{Y_p}} \theta_{M_p} + \beta_{7g}^{\theta_{Y_p}} \theta_{M_p}^2 + \beta_{8g}^{\theta_{Y_p}} \theta_{X_p} \theta_{M_p})}{\# \text{ of iterations}} \quad (51)$$

The difference was summarized as follows:

$$Dif f_{poly_expo}^{\theta_{M_p}} = y_{true_{expo}}^{\theta_{M_p}} - y_{esti_{poly}}^{\theta_{M_p}} \quad (52)$$

$$Dif f_{poly_expo}^{\theta_{Y_p}} = y_{true_{expo}}^{\theta_{Y_p}} - y_{esti_{poly}}^{\theta_{Y_p}}$$

The true sine nonlinear curve was calculated as follows:

$$y_{true_{sine}}^{\theta_{M_p}} = \beta_0^{\theta_{M_p}} + \beta_1^{\theta_{M_p}} \sin(\beta_2^{\theta_{M_p}} \theta_{X_p}) \quad (53)$$

$$y_{true_sine}^{\theta_{Yp}} = \beta_3^{\theta_{Yp}} + \beta_4^{\theta_{Yp}} \sin(\beta_5^{\theta_{Yp}} \theta_{Xp}) + \beta_6^{\theta_{Yp}} \sin(\beta_7^{\theta_{Yp}} \theta_{Mp})$$

The estimated nonlinear curve with a semiparametric Bayesian approach based on a polynomial nonlinear function was calculated as follows:

$$y_{esti_poly}^{\theta_{Mp}} = \frac{\sum_1^{n.iter} \sum_{g=1}^G \pi_g (\beta_{0g}^{\theta_{Mp}} + \beta_{1g}^{\theta_{Mp}} \theta_{Xp} + \beta_{2g}^{\theta_{Mp}} \theta_{Xp}^2)}{\# \text{ of iterations}}$$

$$y_{esti_poly}^{\theta_{Yp}} = \frac{\sum_1^{n.iter} \sum_{g=1}^G \pi_g (\beta_{3g}^{\theta_{Yp}} + \beta_{4g}^{\theta_{Yp}} \theta_{Xp} + \beta_{5g}^{\theta_{Yp}} \theta_{Xp}^2 + \beta_{6g}^{\theta_{Yp}} \theta_{Mp} + \beta_{7g}^{\theta_{Yp}} \theta_{Mp}^2 + \beta_{8g}^{\theta_{Yp}} \theta_{Xp} \theta_{Mp})}{\# \text{ of iterations}} \quad (54)$$

The difference was summarized as follows:

$$Diff_{poly_sine}^{\theta_{Mp}} = y_{true_sine}^{\theta_{Mp}} - y_{esti_poly}^{\theta_{Mp}} \quad (55)$$

$$Diff_{poly_sine}^{\theta_{Yp}} = y_{true_sine}^{\theta_{Yp}} - y_{esti_poly}^{\theta_{Yp}}$$

In study 2, the polynomial nonlinear function was applied to capture the true exponential relationship as well as the sine relationship. A total of 4 chains, 20,000 iterations per chain, 10,000 burn-in, 100 thin were implemented in the analysis of the exponential and sine curves. A total of 400 iterations were used in the analysis for the posterior references.

Chapter 4: Results

Study 1

The mean nonconvergence rate of all parameters in each polynomial, exponential, and sine function across 100 replications were reported in Table 3. The Rhat values of each parameter larger than 1.1 were coded as 1; otherwise, they were coded as 0 in each replication. The mean nonconvergence rate was the average of 1s across 100 replications. The recovery performance between the estimated nonlinear curves and the true nonlinear curves were compared in each nonlinear function with four estimation approaches: parametric, semiparametric at truncation level 5, semiparametric at truncation level 20, and semiparametric at truncation level 200. The range of differences among the four approaches was presented.

Nonconvergence rates. The nonconvergence rate of Study 1 was summarized in Table 3.

Table 3: Mean Nonconvergence Rate across 100 Replications

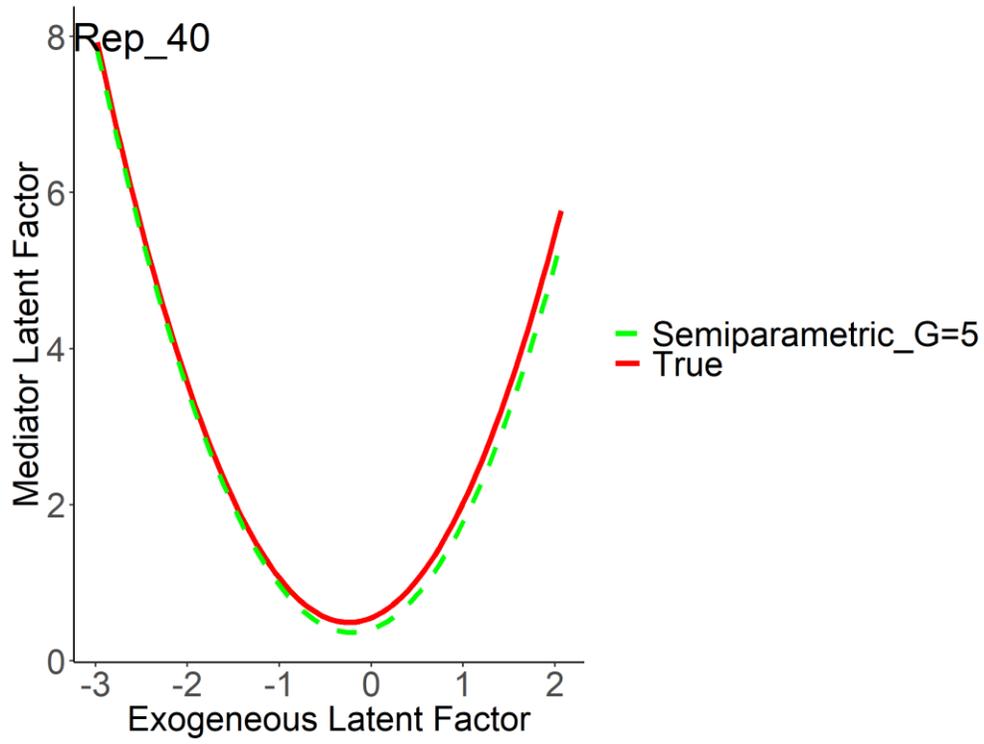
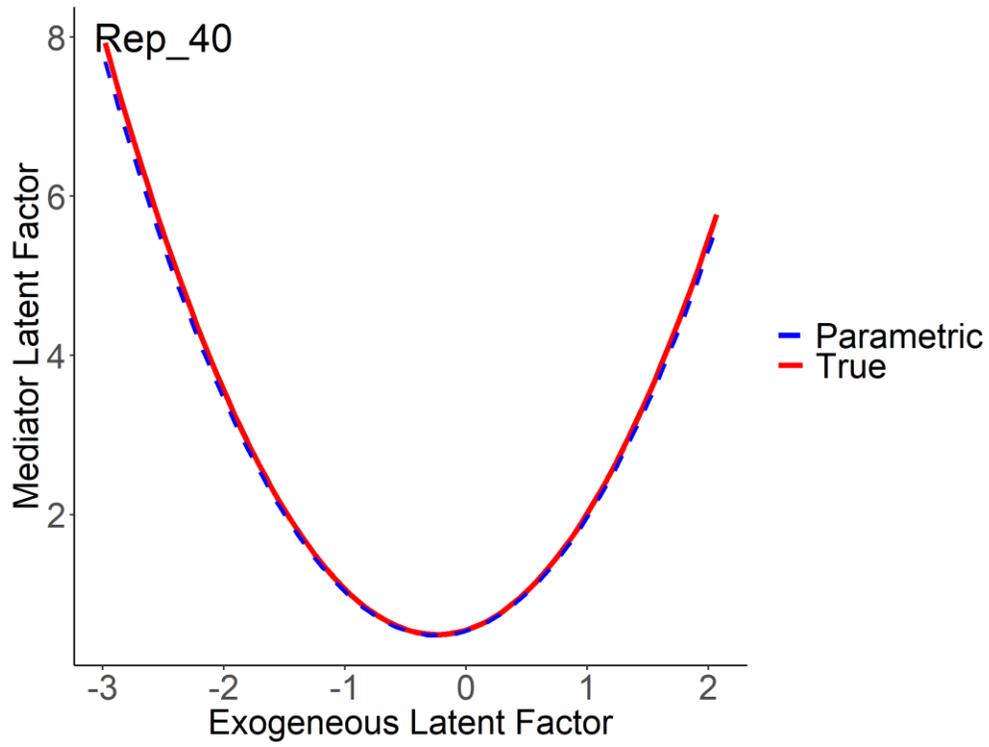
			#	Mean
			parameters	nonconvergence rate
Polynomial	Parametric	G = 1	360	0.002
	Semi-parametric	G = 5	210	0.021
		G = 20	390	0.010
		G = 200	2550	0.004
Exponential	Parametric	G = 1	360	0.000
	Semi-parametric	G = 5	210	0.000
		G = 20	390	0.000
		G = 200	2550	0.000
Sine	Parametric	G = 1	358	0.000
	Semi-parametric	G = 5	205	0.000
		G = 20	370	0.000
		G = 200	2350	0.000

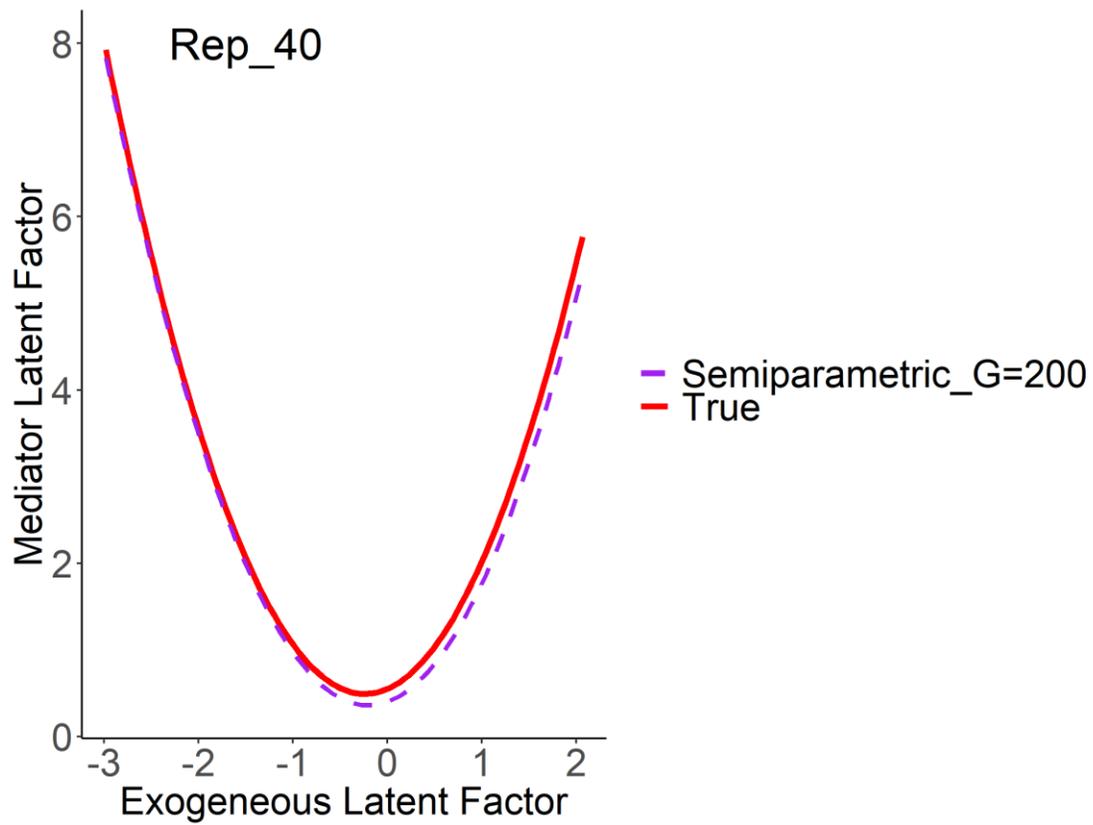
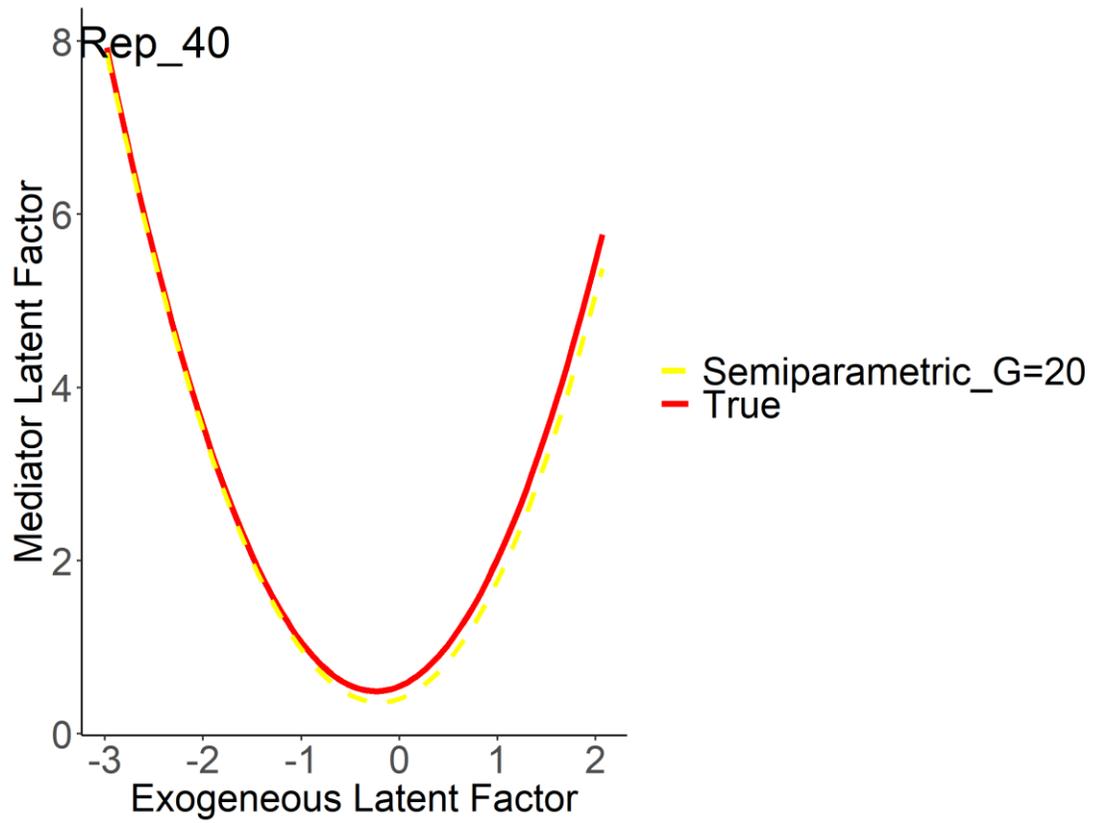
Across 100 replications, 45 replications converged in the polynomial condition with the parametric approach, 33 replications converged in the polynomial condition with the

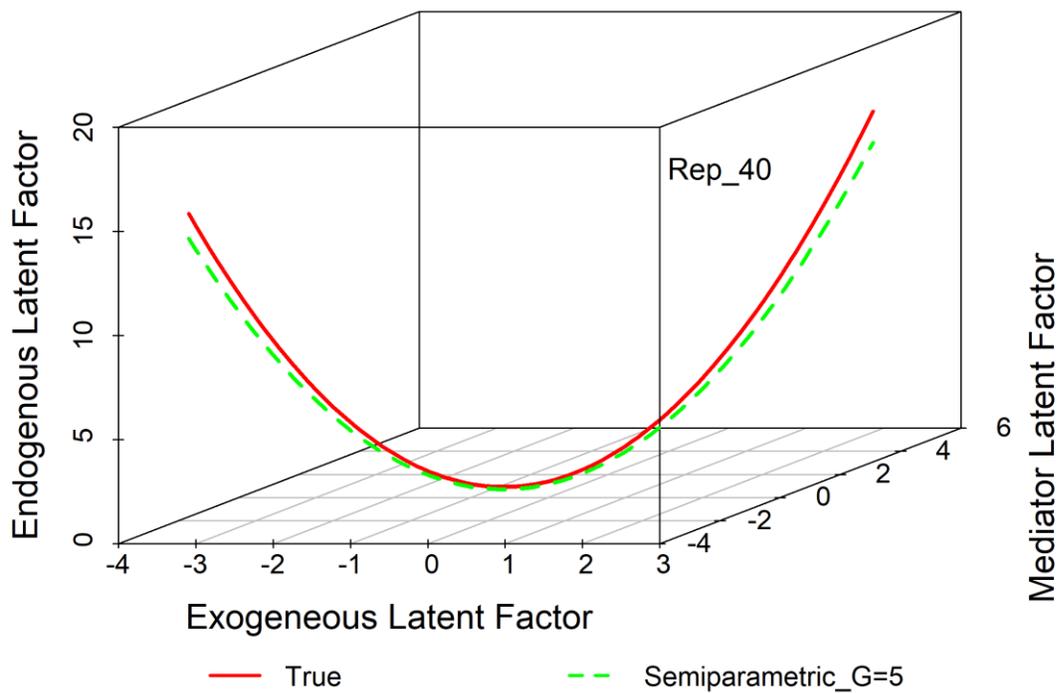
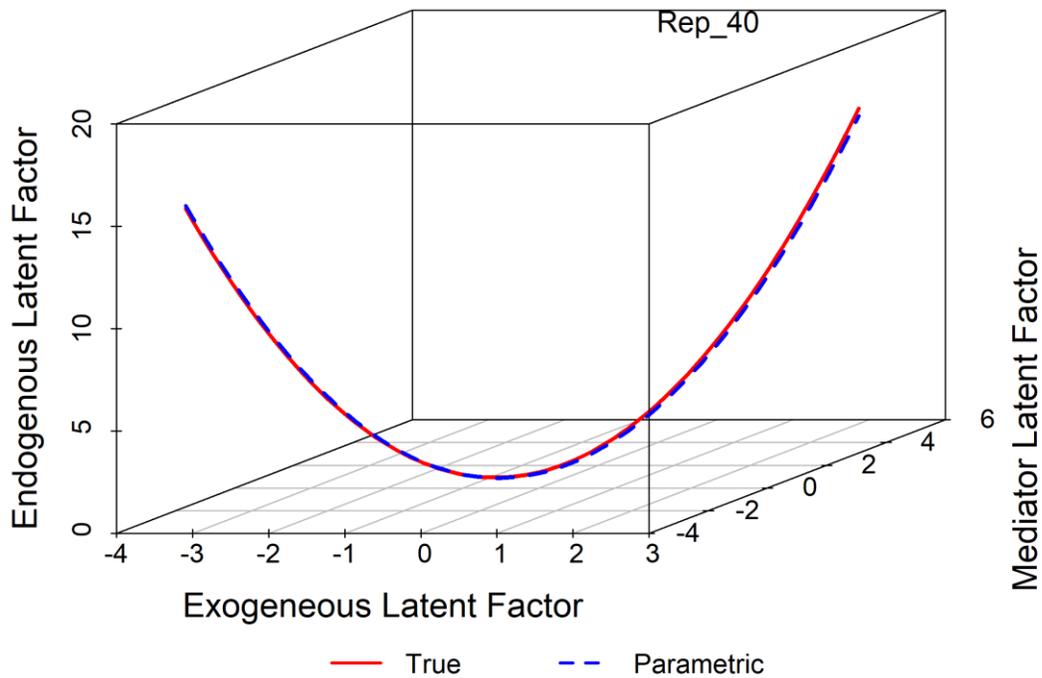
semiparametric approach at truncation level 5, 36 replications converged in the polynomial condition with the semiparametric approach at truncation level 20, and 9 replications converged in the polynomial condition with the semiparametric approach at truncation level 200. The exponential and sine functions had a higher nonconvergence rate across 100 replications. In total, 96 replications converged in the exponential condition with the parametric approach, 96 replications converged in the exponential condition with the semiparametric approach at truncation level 5, 95 replications converged in the exponential condition with the semiparametric approach at truncation level 20, and 69 replications converged in the exponential condition with the semiparametric approach at truncation level 200 across 100 replications. Similarly, across 100 replications, 97 replications converged in the sine condition with the parametric approach, 94 replications converged in the sine condition with the semiparametric approach at truncation level 5, 86 replications converged in the sine condition with the semiparametric approach at truncation level 20, and 58 replications converged in the sine condition with the semiparametric approach at truncation level 200.

Recovery rate. Study 1 tested whether the semiparametric approach more accurately recovered the true nonpolynomial nonlinear curves than the parametric approach. Two nonlinear direct relations were investigated: (a) the nonlinear relations between the exogenous latent predictor (θ_X) and the mediator latent factor (θ_M) ($\theta_M - \theta_X$), and (b) the nonlinear relations between the endogenous latent factor (θ_Y) and two exogenous latent predictors (θ_X, θ_M) ($\theta_Y - \theta_M \theta_X$).

Polynomial nonlinear function. As shown in Figure 5, the semiparametric approach and the parametric approach had similar recoveries in estimating the $\theta_M - \theta_X$ and $\theta_Y - \theta_M \theta_X$ polynomial relationship.







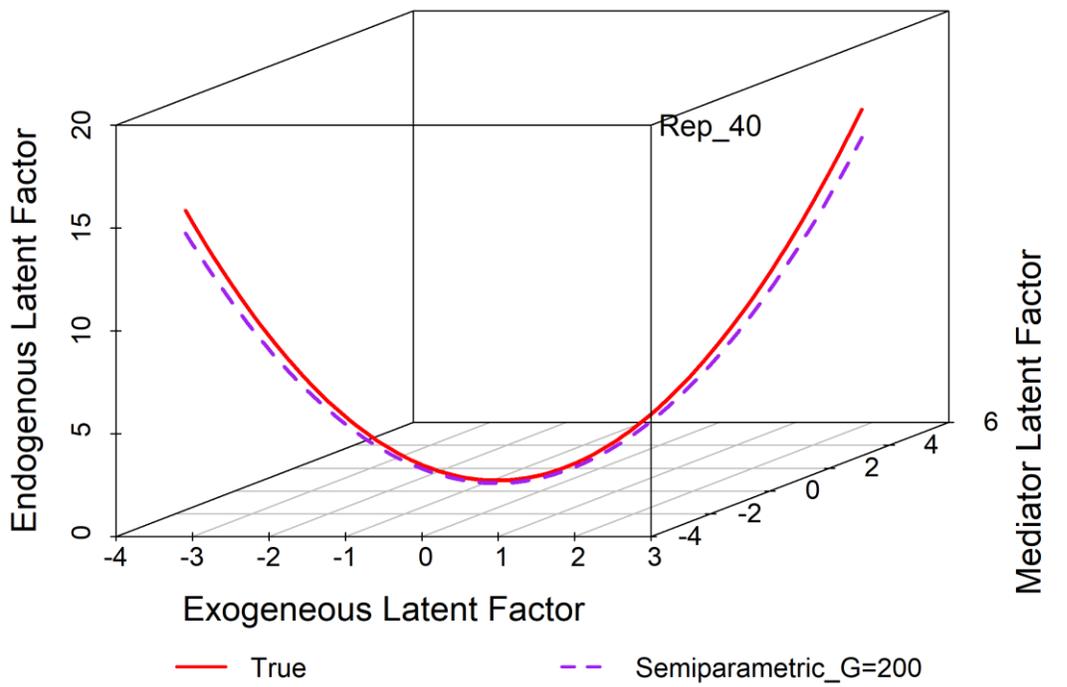
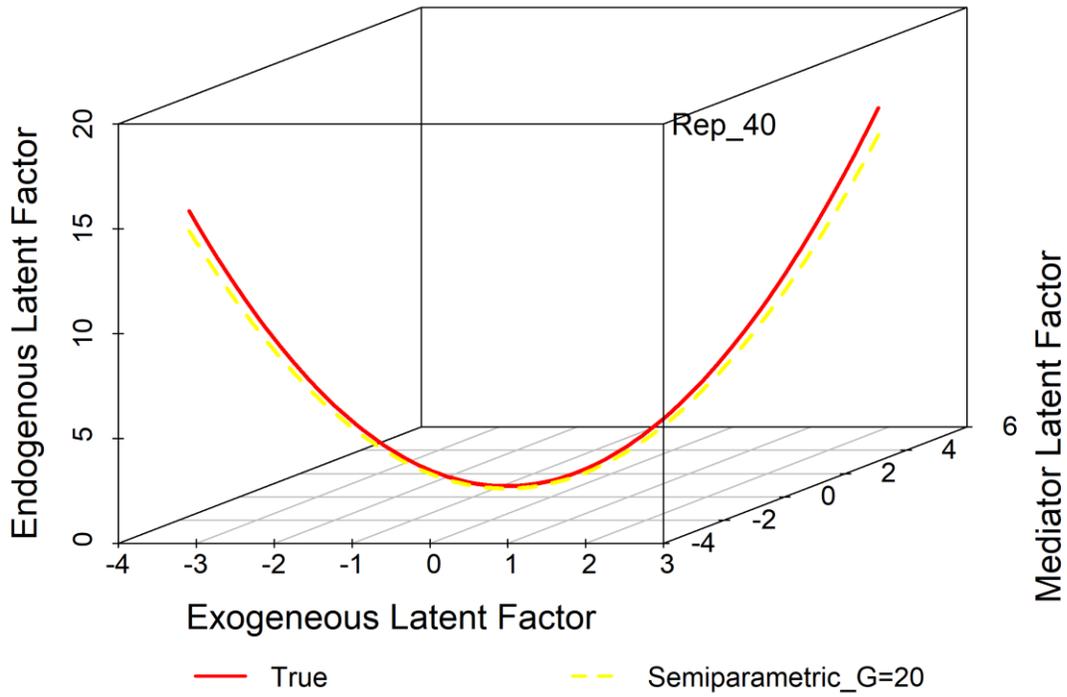
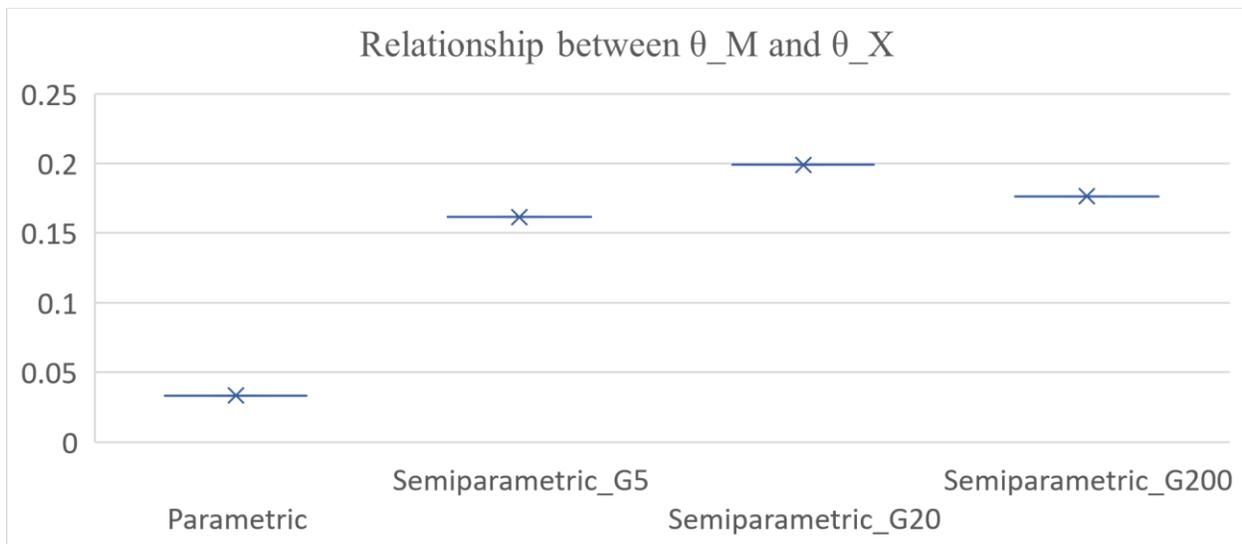


Figure 5: The $\theta_M - \theta_X$ and $\theta_Y - \theta_M \theta_X$ polynomial curves estimated by the parametric and the semiparametric approach

The semiparametric approach had a larger range of differences between the true polynomial curve and the estimated polynomial curve than the parametric approach. The range of differences varied from 0 to 0.4 in the estimation of the $\theta_M - \theta_X$ relationship and from 0 to 6.25 in the estimation of the $\theta_Y - \theta_M \theta_X$ relationship with the semiparametric approach. In contrast, the range of differences varied from 0 to 0.3 in the estimation of the $\theta_M - \theta_X$ direct effect and between 0 and 2 in the estimation of the $\theta_Y - \theta_M \theta_X$ direct effect with the parametric approach.



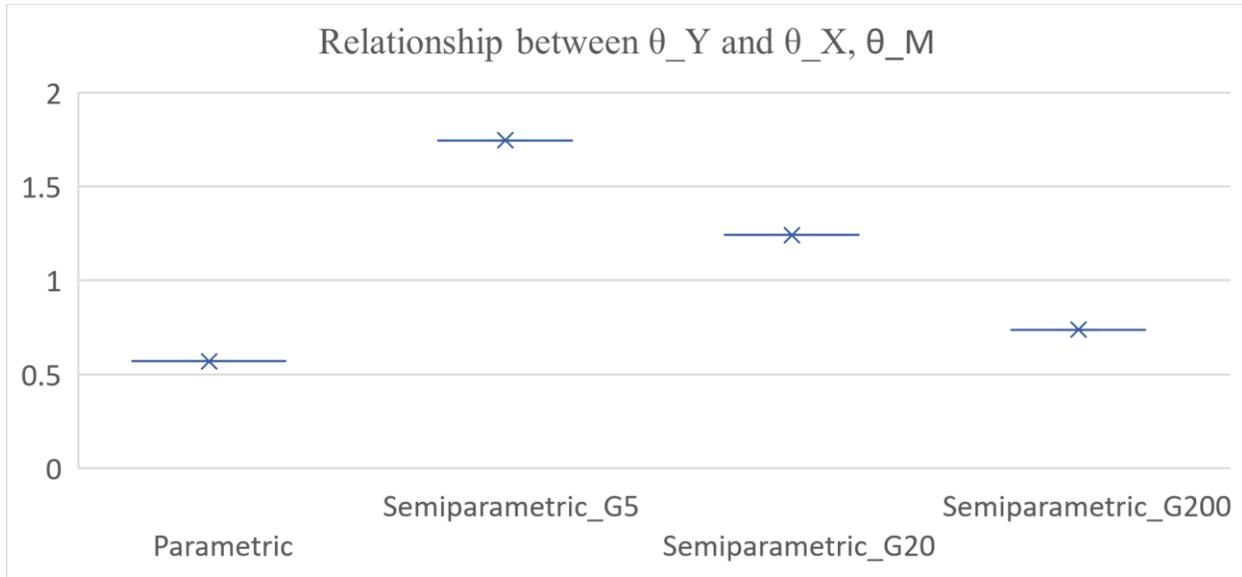
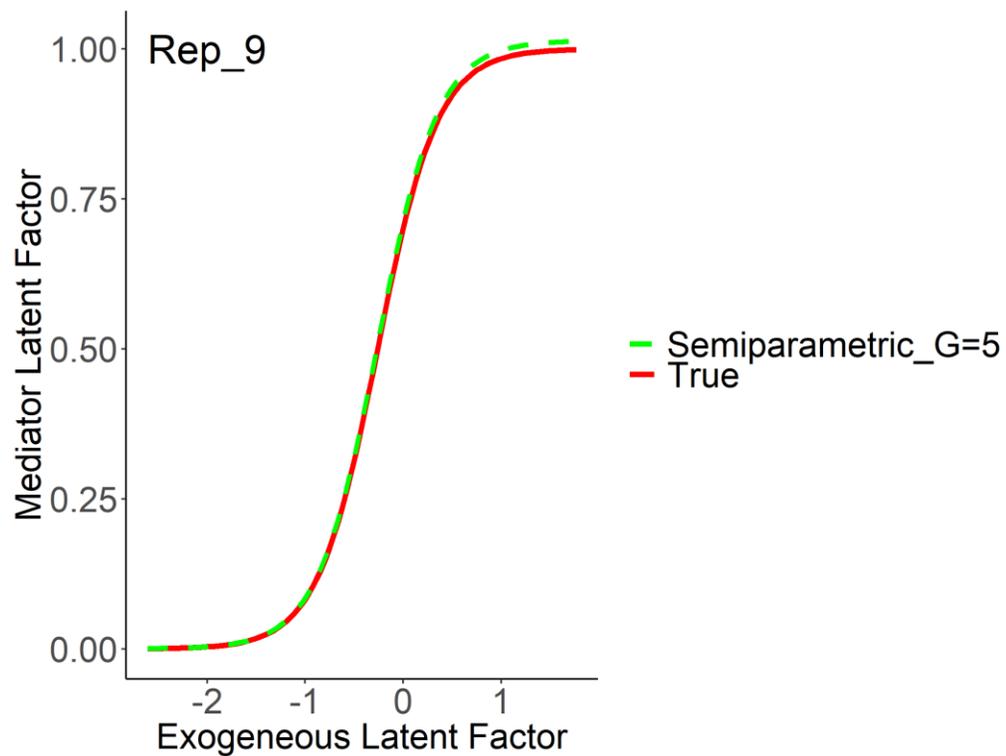
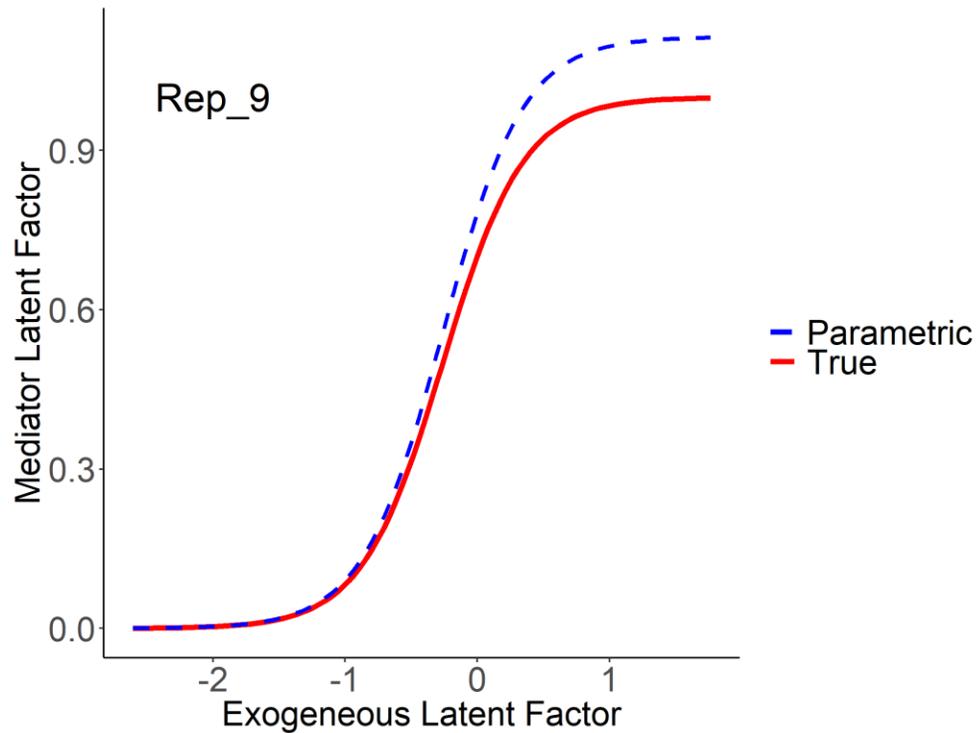


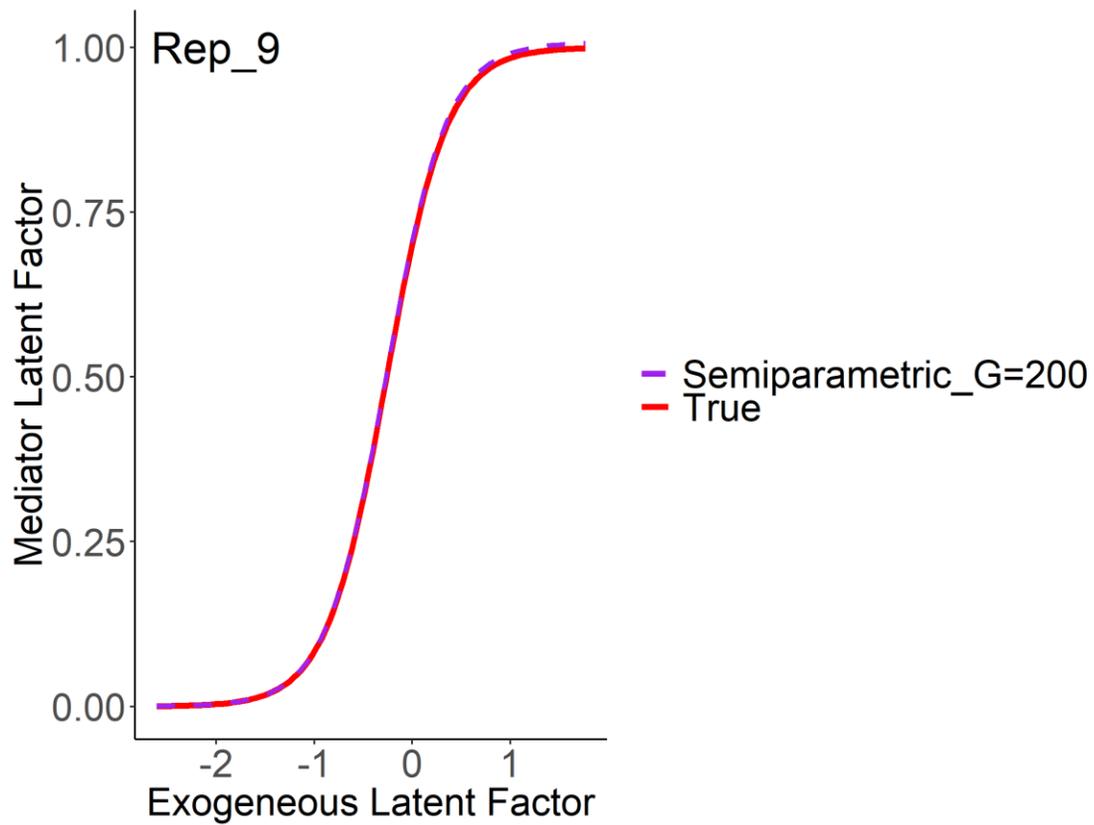
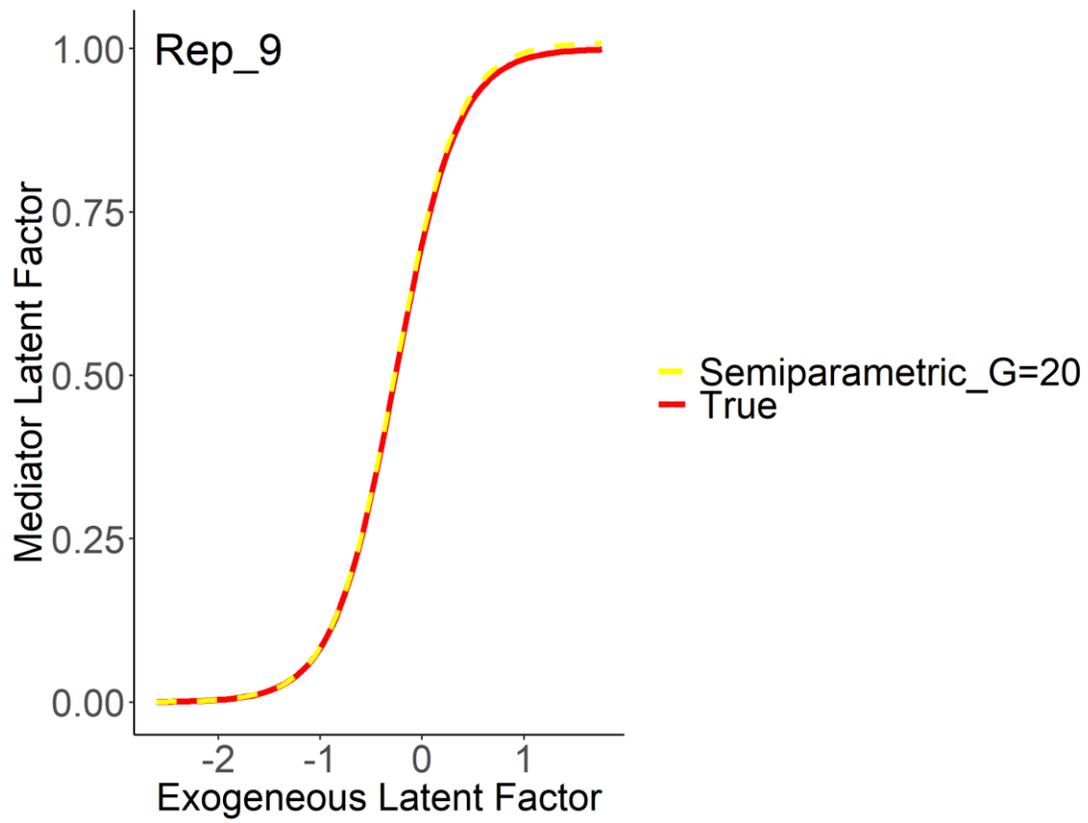
Figure 6. The mean range of differences in the $\theta_M - \theta_X$ and $\theta_Y - \theta_M \theta_X$ polynomial curves

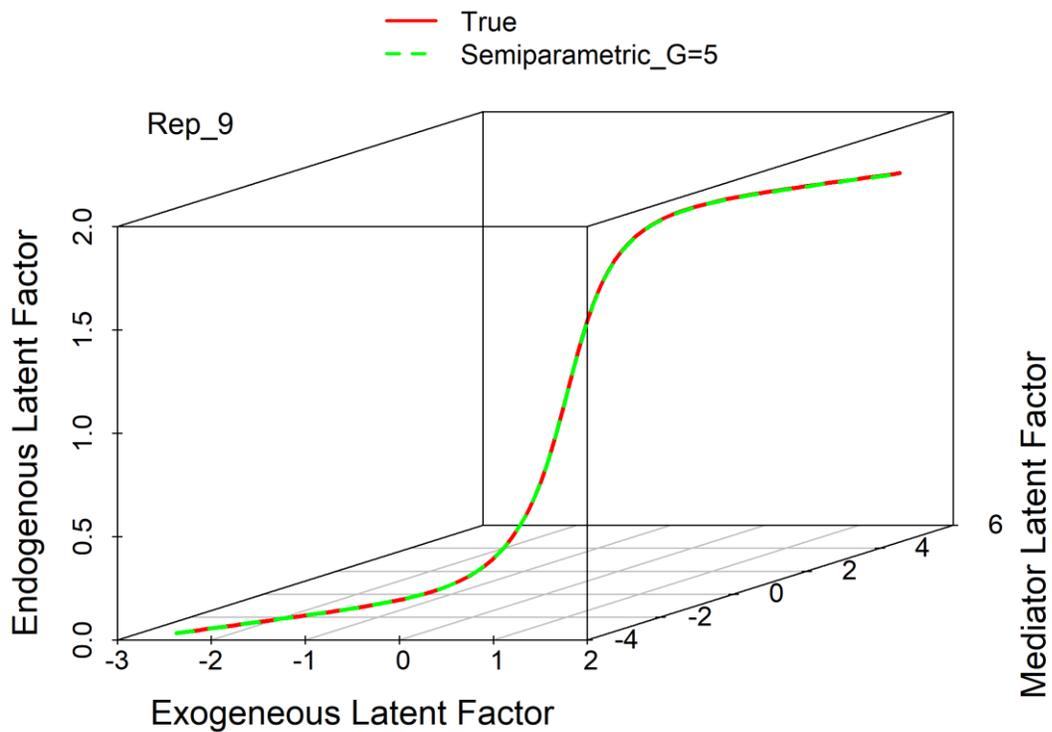
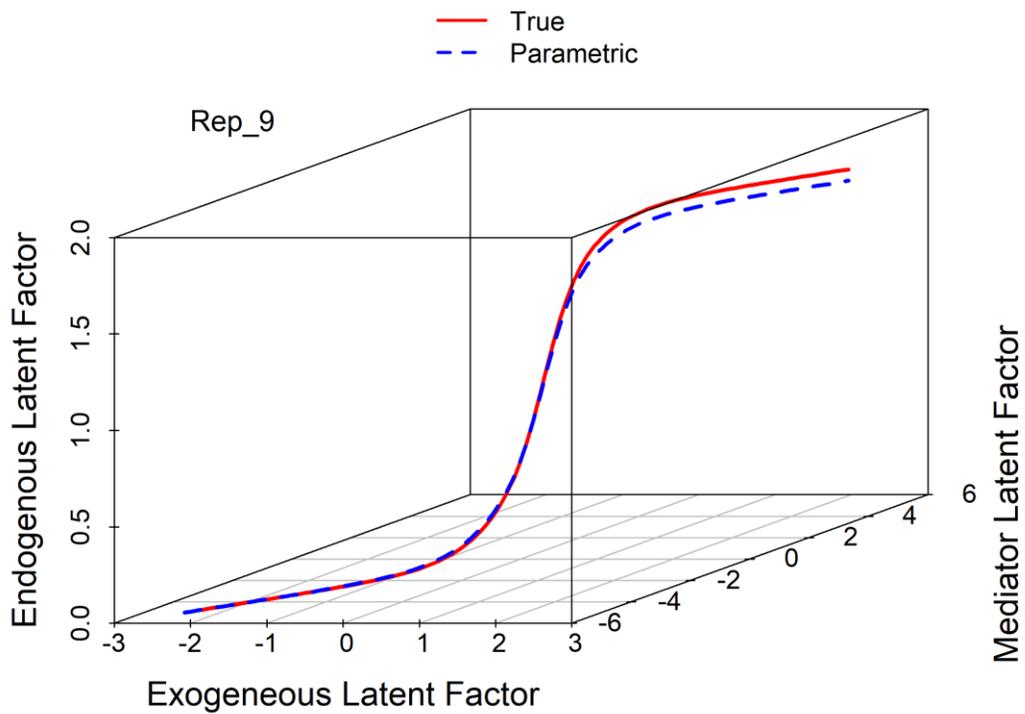
A t-test was used to compare the mean of difference range between the true polynomial curves and the estimated polynomial curves across parametric and semiparametric approach. The results found the parametric approach had a significantly higher accuracy than the semiparametric approach at truncation level 5 and 20 ($t=0.08$, $p<.001$; $t=0.12$, $p<.001$) in capturing the true $\theta_M - \theta_X$ polynomial relationship. However, the difference between the parametric approach and the semiparametric approach at truncation level 200 was not significant. In addition, the semiparametric approach at truncation level 200 was significantly better than the semiparametric approach at truncation level 5 and 20 ($t=-0.08$, $p<.001$; $t=-0.12$, $p<.001$) in recovering the $\theta_M - \theta_X$ polynomial relationship.

Similar results were found in recovering the $\theta_Y - \theta_M \theta_X$ polynomial relationship. The semiparametric approach at truncation level 200 was significantly better than the semiparametric approach at truncation level 5 and 20 ($t=-1.13$, $p<.001$; $t=-0.84$, $p<.001$). However, the parametric approach was only significantly better than the semiparametric approach at truncation level 5 ($t=0.70$, $p<.05$).

Exponential nonlinear function. However, the semiparametric approach more accurately captured the $\theta_M - \theta_X$ and $\theta_Y - \theta_M \theta_X$ exponential relationship than the parametric approach.







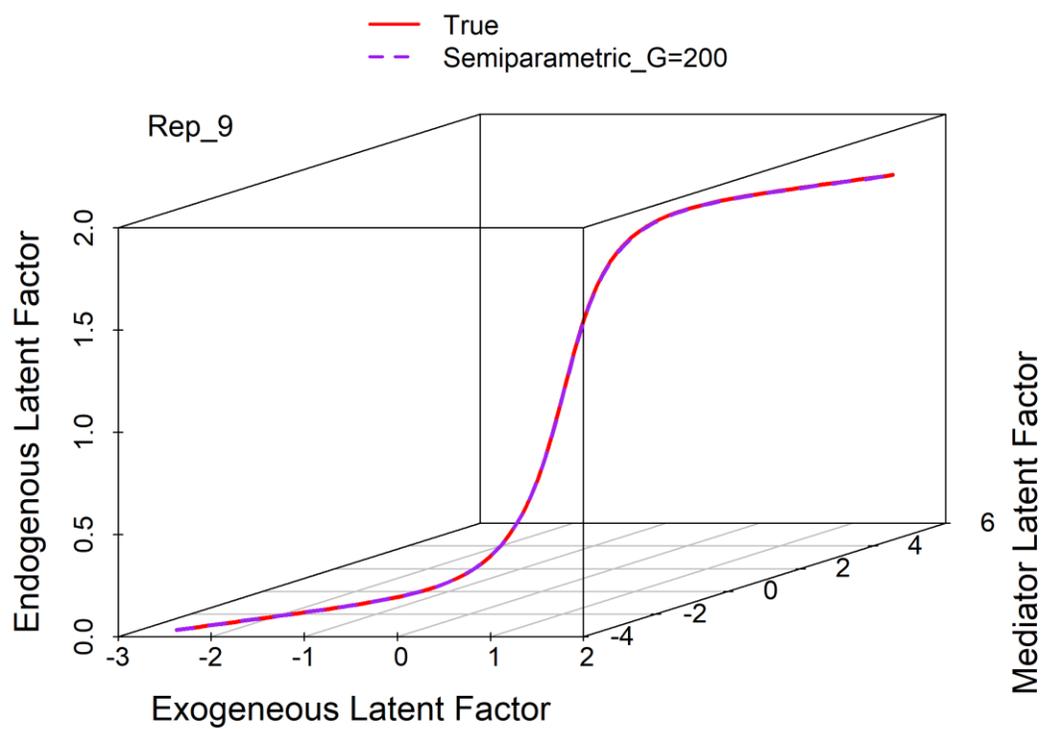
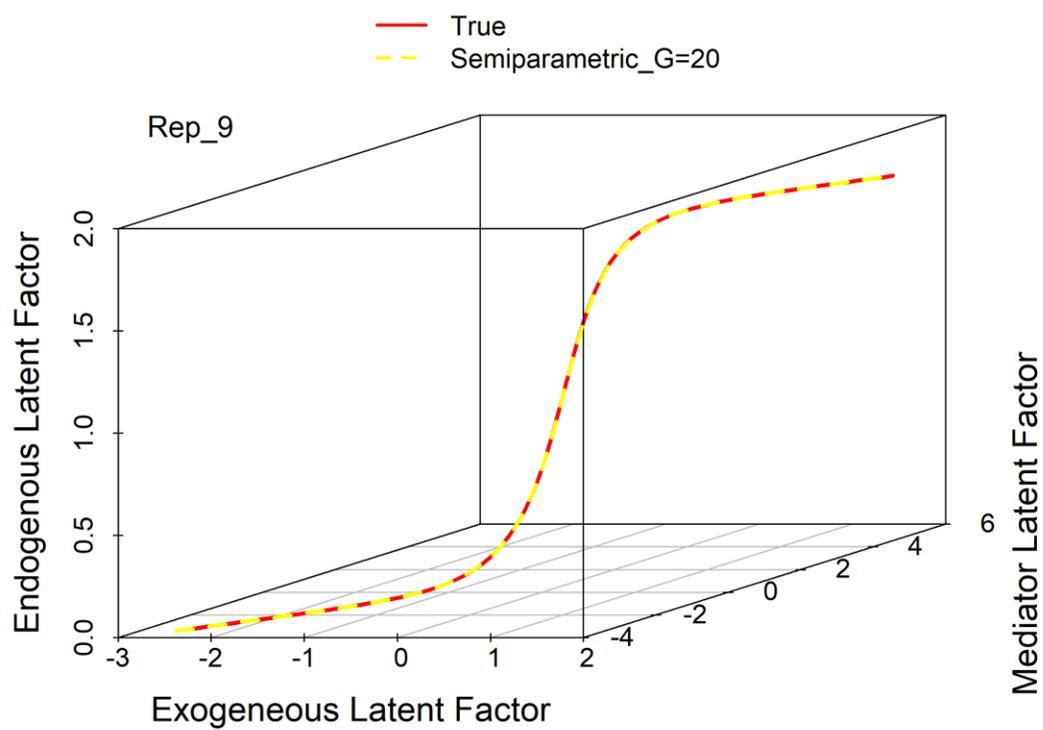
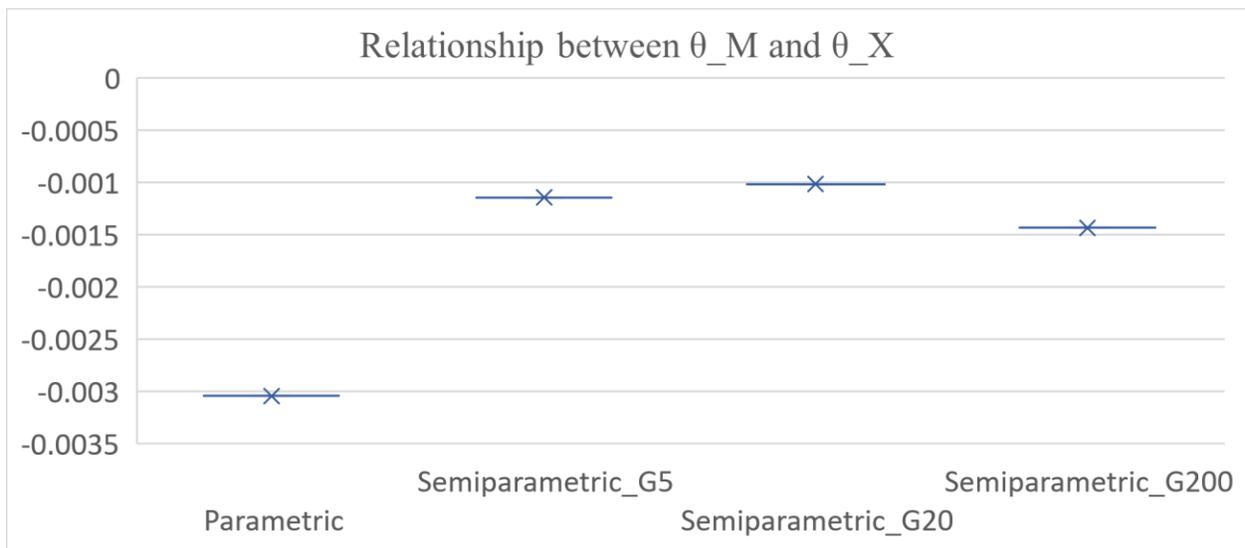


Figure 7: The $\theta_M - \theta_X$ and $\theta_Y - \theta_M \theta_X$ exponential curves estimated by the parametric and semiparametric approaches

The parametric approach had a larger range of differences between the true exponential curve and the estimated exponential curve than the semiparametric approach. The range of differences varied from 0 to 0.05 in the estimation of the $\theta_M - \theta_X$ relationship and from 0 to 0.1 in the estimation of the $\theta_Y - \theta_M \theta_X$ relationship with the parametric approach. In contrast, the range of differences varied from 0 to 0.03 in the estimation of the $\theta_M - \theta_X$ and $\theta_Y - \theta_M \theta_X$ relationships with the semiparametric approach. However, the differences between the parametric approach and the semiparametric approach were not significant.



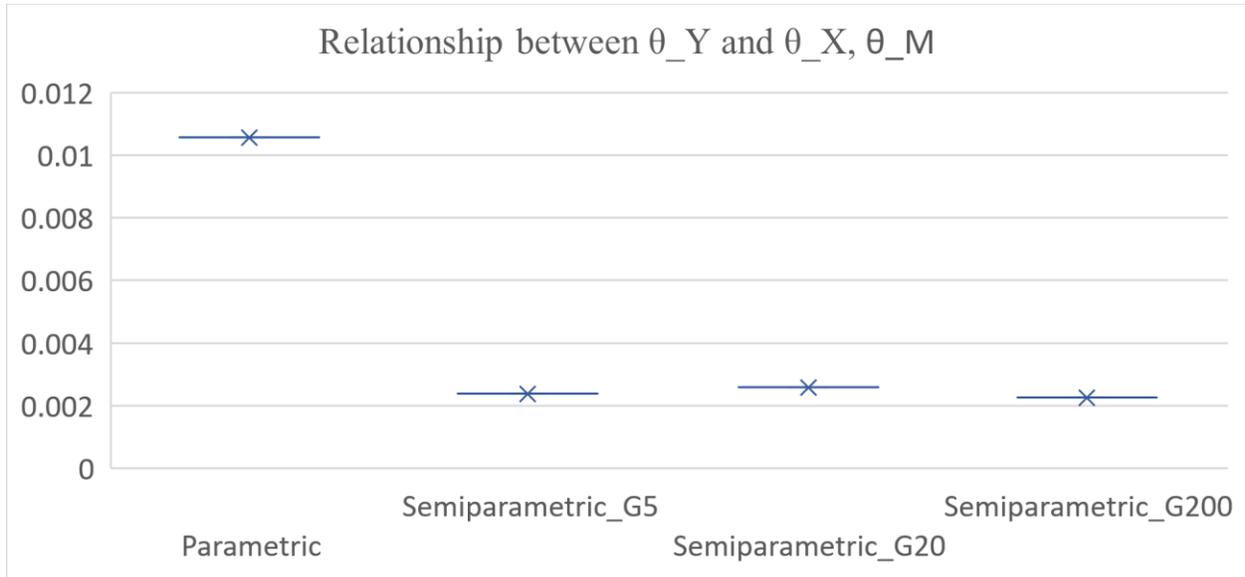
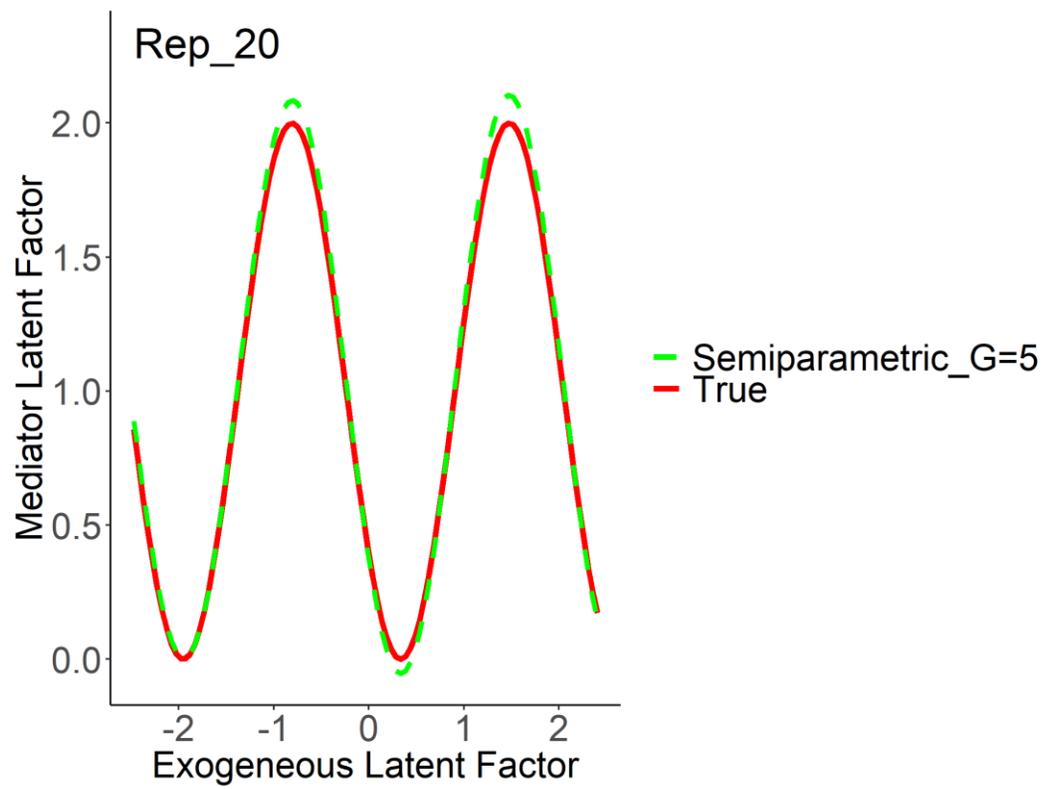
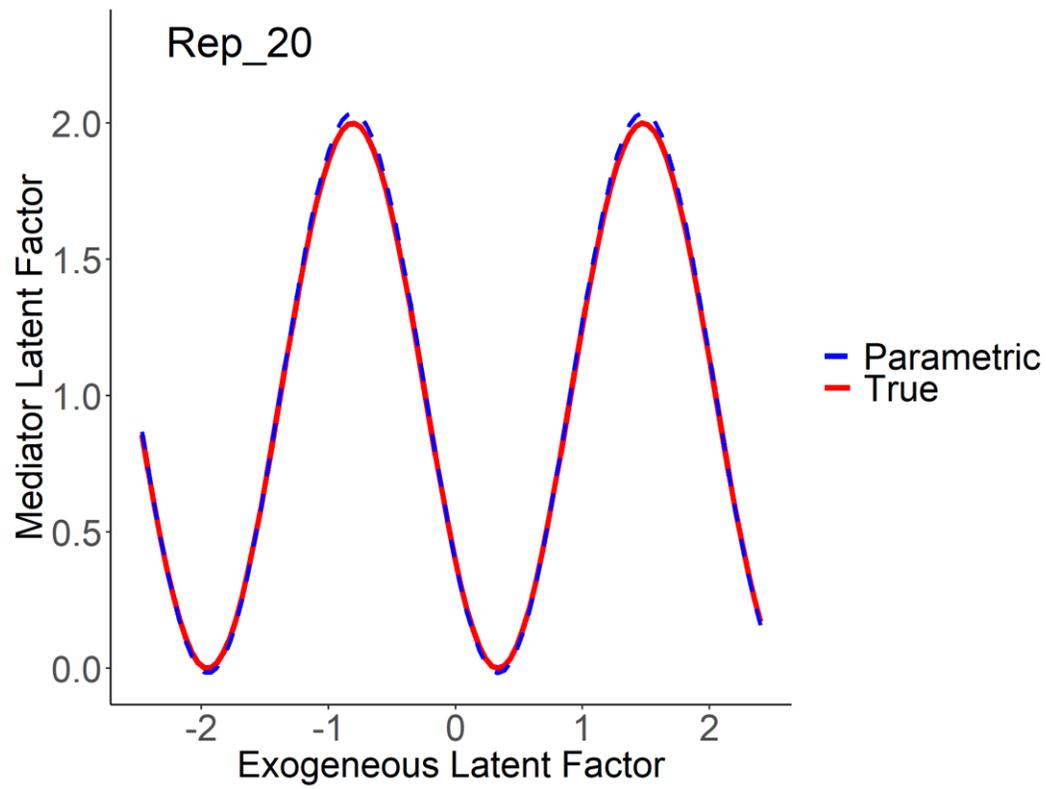
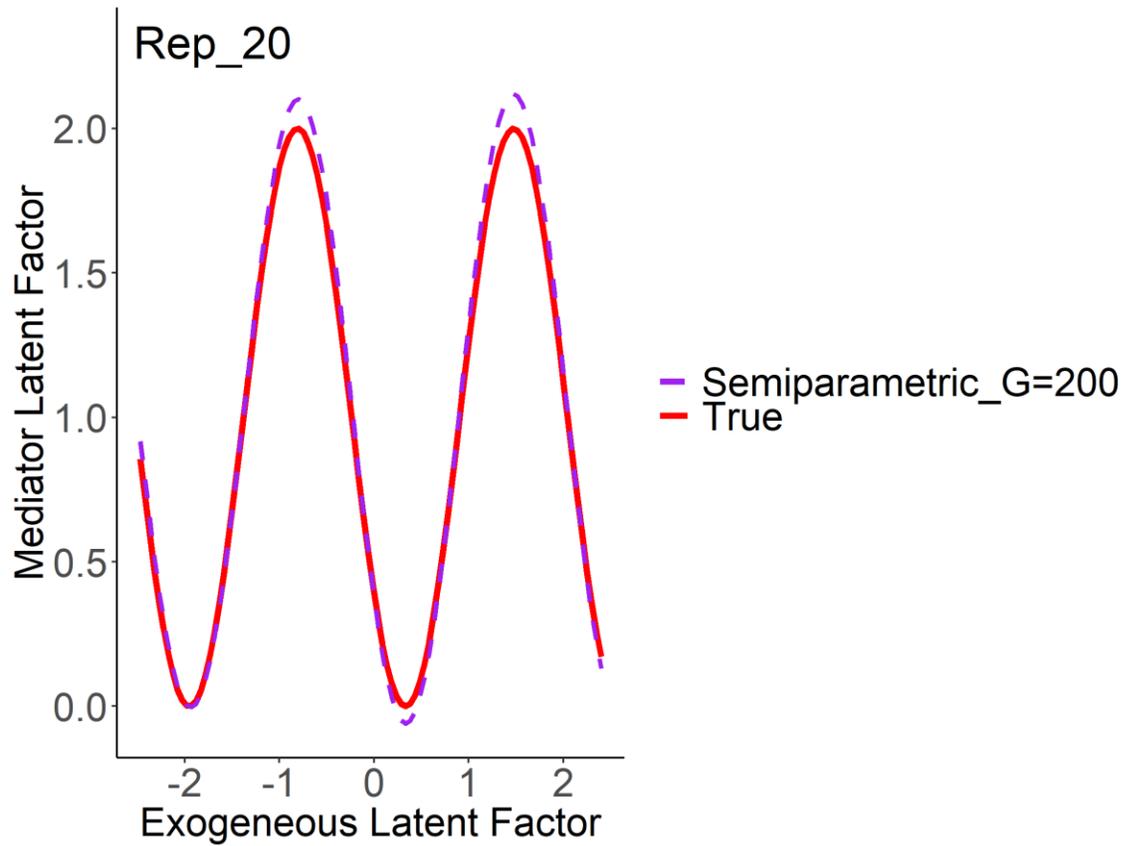
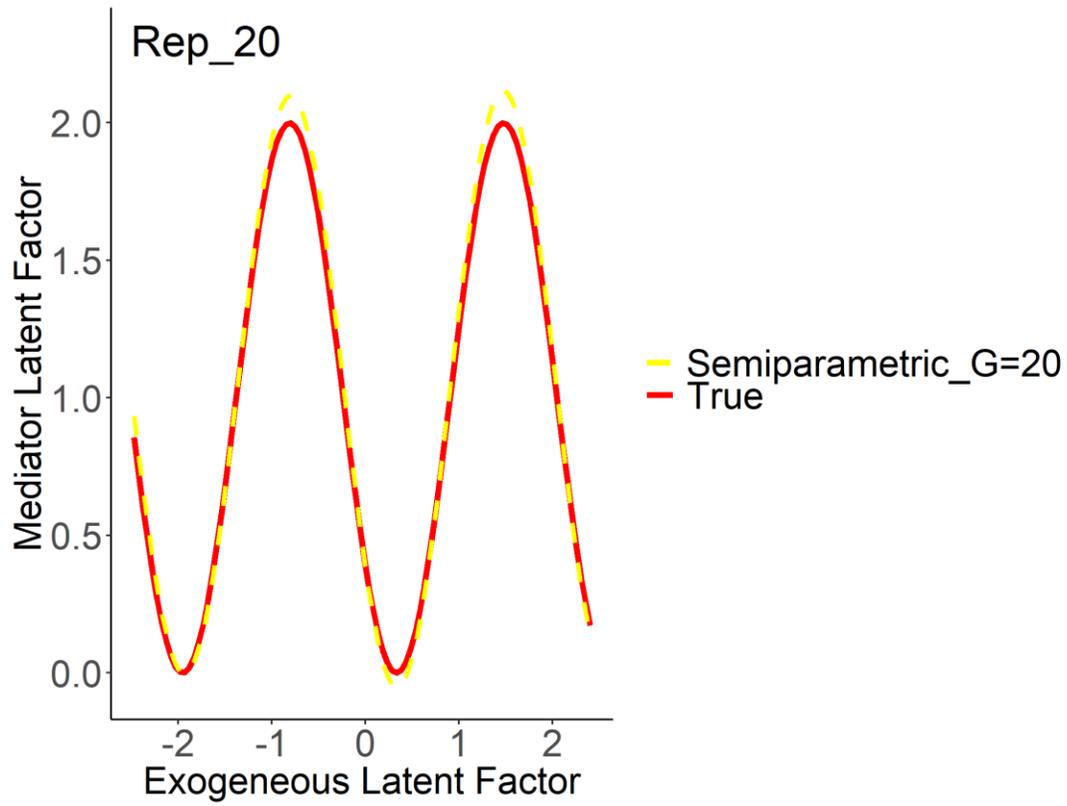
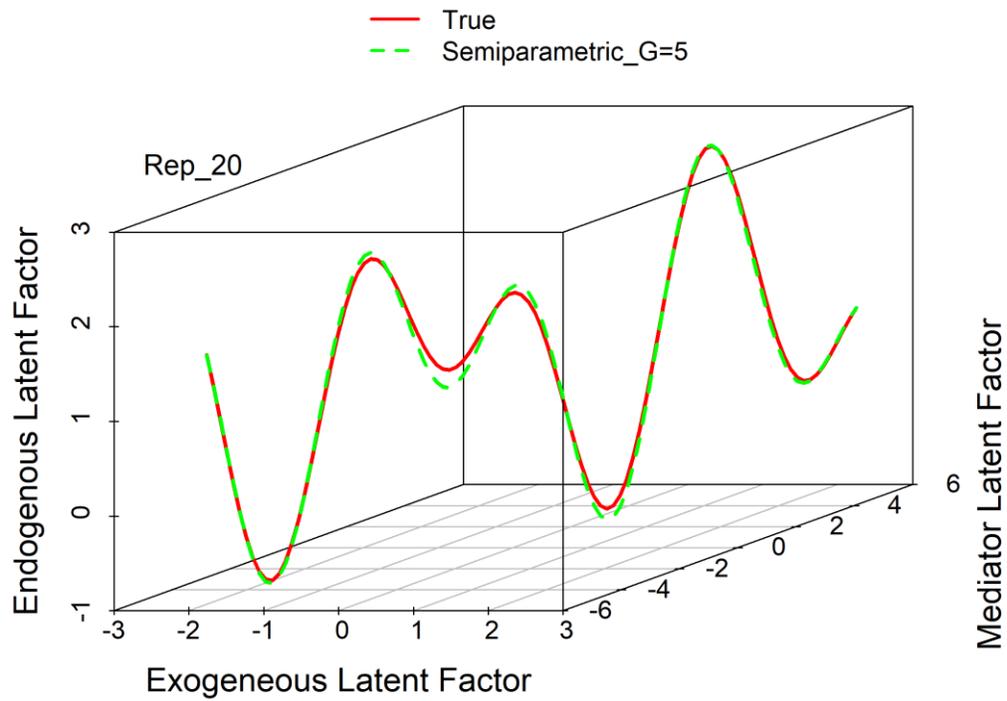
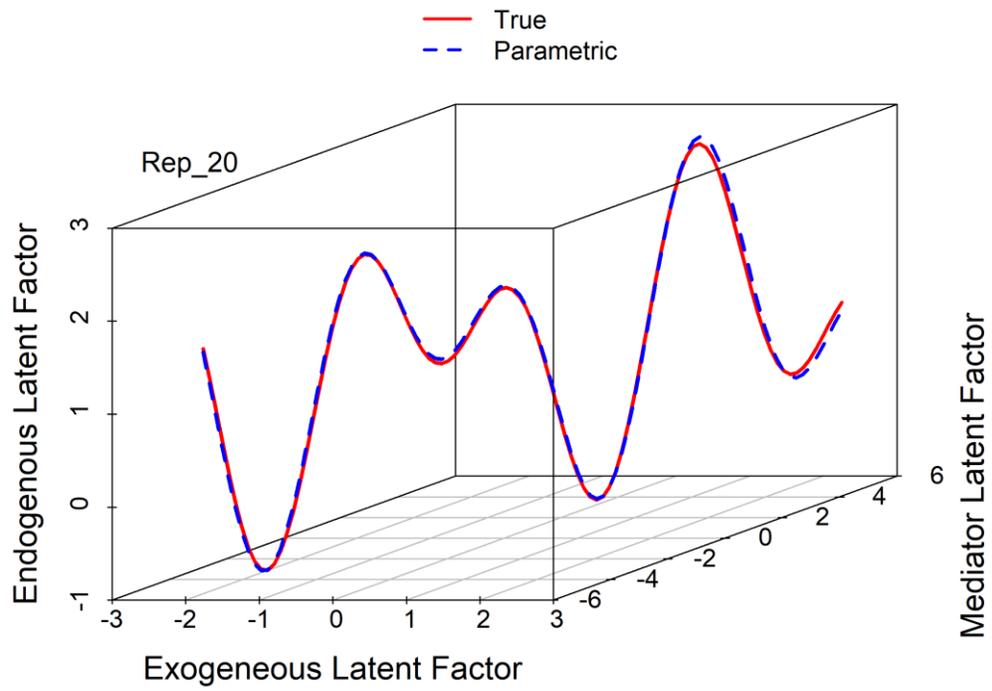


Figure 8: The mean range of differences in the $\theta_M - \theta_X$ and $\theta_Y - \theta_M \theta_X$ exponential curves

Sine nonlinear function. The semiparametric approach performed similarly in recovering the $\theta_M - \theta_X$ and $\theta_Y - \theta_M \theta_X$ sine nonlinear curves as the parametric approach, as shown in Figure 9.







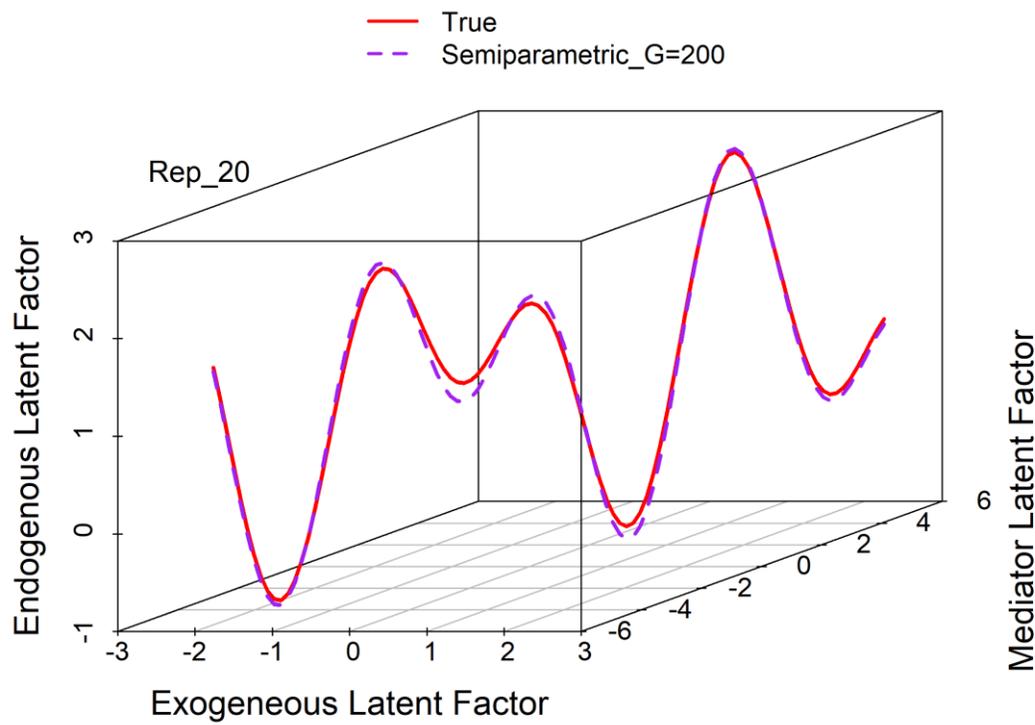
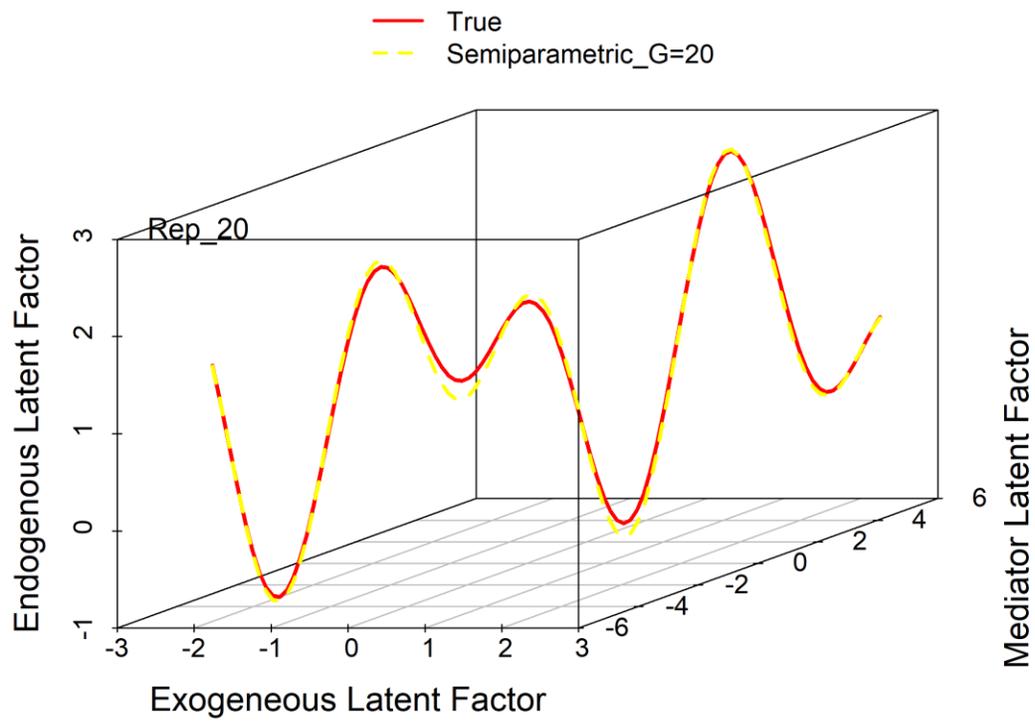


Figure 9: The $\theta_M - \theta_X$ and $\theta_Y - \theta_M \theta_X$ sine curves estimated by the parametric and semiparametric approaches

The range of difference between the true sine curve and the estimated sine curve in the estimation of the $\theta_M - \theta_X$ and $\theta_Y - \theta_M \theta_X$ relationships with the semiparametric approach ranged from 0 to 0.15, and the range of difference between the true sine curve and the estimated curve in the estimation of the $\theta_M - \theta_X$ and $\theta_Y - \theta_M \theta_X$ relationships with the parametric approach ranges from 0 to 0.1. Similarly, the differences between the semiparametric approach and parametric approach in estimating the $\theta_M - \theta_X$ and $\theta_Y - \theta_M \theta_X$ sine curves were not significant.

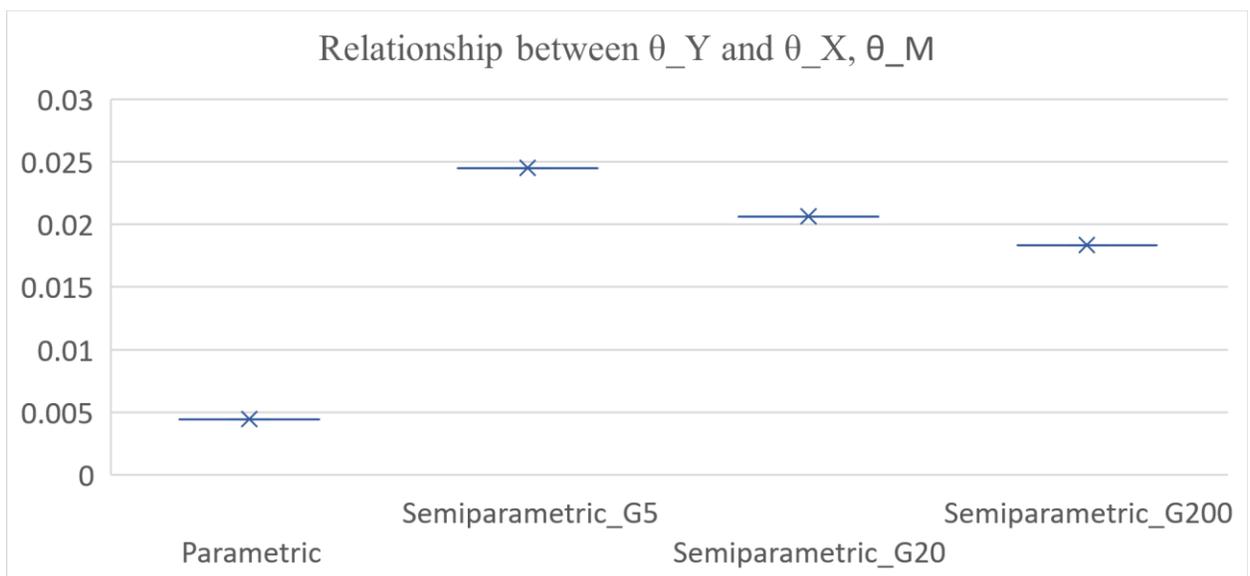
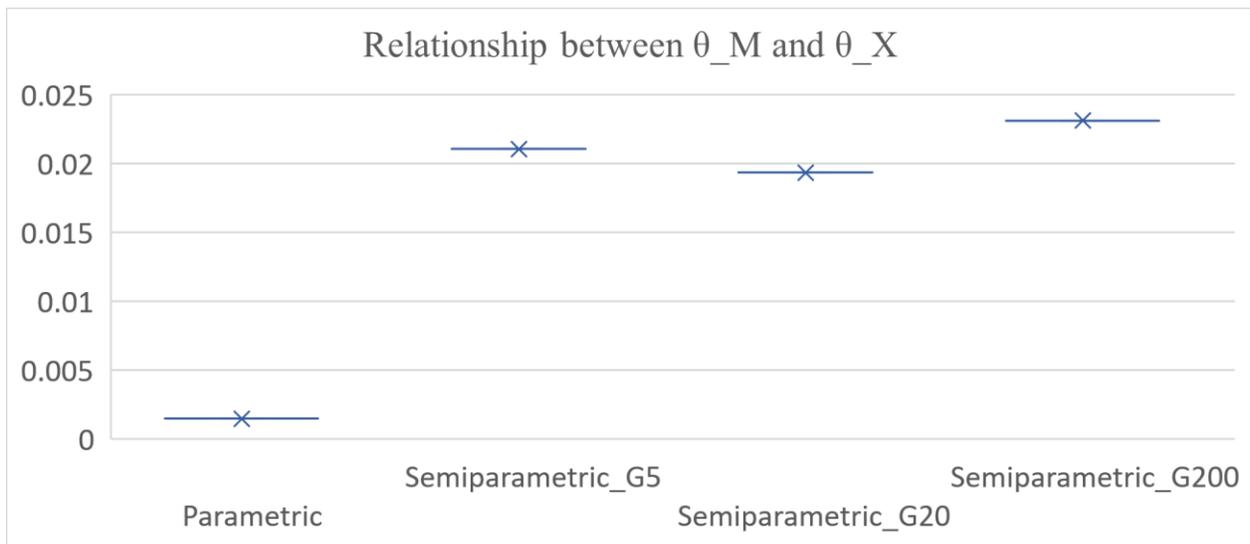


Figure 10: The mean range of differences in the $\theta_M - \theta_X$ and $\theta_Y - \theta_M \theta_X$ sine curves

In summary, the parametric approach was significantly better in recovering the $\theta_M - \theta_X$ and $\theta_Y - \theta_M \theta_X$ polynomial relationships than the semiparametric approach at truncation level 5 and 20. The semiparametric approach at truncation level 200 performed similarly to the parametric approach in recovering the $\theta_M - \theta_X$ and $\theta_Y - \theta_M \theta_X$ polynomial relationships. However, the semiparametric approach had a higher accuracy in capturing the $\theta_M - \theta_X$ and $\theta_Y - \theta_M \theta_X$ exponential and sine relationships than the parametric approach, in which there were no differences among truncation levels within the semiparametric approach. The truncation level at 5 was sufficient to capture the exponential nonlinearity.

Study 2

Study 2 explored whether the semiparametric approach captures the true nonpolynomial relations when the polynomial function was pre-assumed. Specifically, the polynomial function with the quadratic and interaction effect was applied in the structural model to recover the true exponential and sine curves with the Bayesian semiparametric approach.

Nonconvergence rate. Due to the capacity of computer memory and running time, 50 replications were ran in the study 2. The mean nonconvergence rates of all parameters in the exponential and sine functions across 50 replications were reported in Table 4.

Table 4: Mean Nonconvergence Rate across 50 Replications

		# parameters	Mean nonconvergence rate
Poly_ Exponential	$G = 5$	211	0.000
	$G = 20$	391	0.049
	$G = 200$	2551	0.052
Poly_Sine	$G = 5$	211	0.000
	$G = 20$	391	0.049
	$G = 200$	2551	0.043

Across 50 replications, 41 replications converged in the exponential condition with the semiparametric approach at truncation level 5, 5 replications converged in the exponential condition with the semiparametric approach at truncation level 20, and 0 replications converged in the exponential condition with the semiparametric approach at truncation level 200. In addition, 47 replications converged in the sine condition with the semiparametric approach at truncation level 5, 16 replications converged in the sine condition with the semiparametric approach at truncation level 20, and 0 replications converged in the sine condition with the semiparametric approach at truncation level 200.

Recovery rate. The plots in Figure 11 were drawn based on the converged replications at truncation level 5. Neither the exponential or sine $\theta_M - \theta_X$ and $\theta_Y - \theta_M \theta_X$ nonlinearity were captured by the prespecified polynomial function with the semiparametric approach.

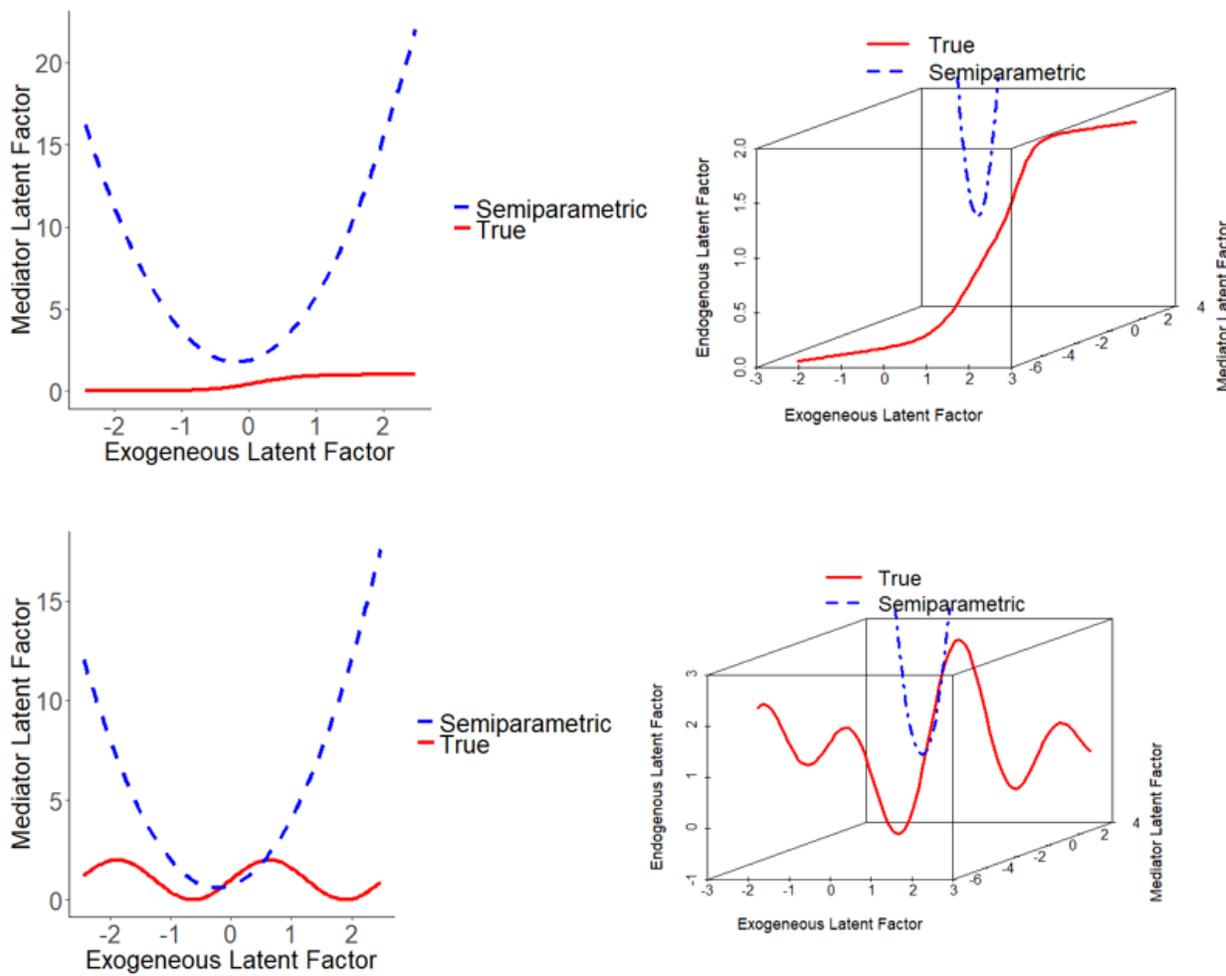


Figure 11: The exponential and sine curves estimated by the semiparametric approach with the polynomial nonlinear function

The significant differences between the estimated nonlinear curve and the true nonlinear curve indicated the importance of specifying a correct functional form when capturing the nonlinear relationships.

Chapter 5: Discussion

This dissertation first compared the parametric approach and the semiparametric approach in capturing the polynomial and nonpolynomial relationships among latent factors in the structural model and then investigated the performance of the semiparametric approach at low, medium, and large truncation levels ($G=5, 20, 200$) in recovering the nonpolynomial relationships when the functional form was misspecified as the polynomial function. The objective of this dissertation was to recover the nonlinear relationships among latent factors in the structural model under different conditions. More specifically, it addressed the following research questions:

1. What are the differences between the true polynomial nonlinear curve and the estimated nonlinear curve with the parametric Bayesian approach?
2. What are the differences between the true polynomial nonlinear curve and the estimated nonlinear curve with the semiparametric Bayesian approach when latent groups of DP prior are small, medium, and large?
3. What are the differences between the true exponential nonlinear curve and the estimated nonlinear curve with the parametric Bayesian approach?
4. What are the differences between the true exponential nonlinear curve and the estimated nonlinear curve with the semiparametric Bayesian approach when latent groups of DP prior are small, medium, and large?
5. What are the differences between the true sinusoidal nonlinear curve and the estimated nonlinear curve with the parametric Bayesian approach?

6. What are the differences between the true sinusoidal nonlinear curve and the estimated nonlinear curve with the semiparametric Bayesian approach when latent groups of DP prior are small, medium, and large?
7. How well is the true exponential nonlinear curve recovered by the polynomial nonlinear functions with the semiparametric Bayesian approach when latent groups of DP prior are small, medium, and large?
8. How well is the true sinusoidal nonlinear curve recovered by the polynomial nonlinear functions with the semiparametric Bayesian approach when latent groups of DP prior are small, medium, and large?

To answer the research questions listed above, one simulation study was conducted. Two analyses were developed in the simulation study to evaluate (1) two proposed estimation approaches in terms of nonlinearity recoveries under different combinations of two design factors (the type of nonlinear functions (polynomial, exponential, and sine) and the truncation levels (1, 5, 20, 200)) and (2) the proposed semiparametric approach in terms of its nonpolynomial nonlinearity recoveries when the nonlinear function was misspecified under different combinations of conditions. More specifically, the simulation study has found:

1. The parametric approach had a smaller range of differences between the true polynomial curve and the estimated polynomial curve than the semiparametric approach. The range of differences varied from 0 to 0.3 in the estimation of the $\theta_M - \theta_X$ direct effect and between 0 and 2 in the estimation of the $\theta_Y - \theta_M \theta_X$ direct effect with the parametric approach.
2. The semiparametric approach had a larger range of differences between the true polynomial curve and the estimated polynomial curve than the parametric approach. The

range of differences varied from 0 to 0.4 in the estimation of the $\theta_M - \theta_X$ relationship and from 0 to 6.25 in the estimation of the $\theta_Y - \theta_M \theta_X$ relationship with the semiparametric approach. In addition, the semiparametric approach at truncation level 200 was significantly better than the semiparametric approach at truncation level 5 and 20 in recovering the $\theta_M - \theta_X$ and $\theta_Y - \theta_M \theta_X$ polynomial relationship.

3. The range of differences between the true exponential curve and the estimated exponential curve varied from 0 to 0.03 in the estimation of the $\theta_M - \theta_X$ and $\theta_Y - \theta_M \theta_X$ relationships with the parametric approach.
4. The range of differences between the true exponential curve and the estimated exponential curve varied from 0 to 0.05 in the estimation of the $\theta_M - \theta_X$ relationship and from 0 to 0.1 in the estimation of the $\theta_Y - \theta_M \theta_X$ relationship with the semiparametric approach. There was no significant difference among truncation levels. The truncation level at 5 was sufficient to capture the exponential relationship.
5. The range of difference between the true sine curve and the estimated curve in the estimation of the $\theta_M - \theta_X$ and $\theta_Y - \theta_M \theta_X$ relationships with the parametric approach ranged from 0 to 0.1.
6. The range of difference between the true sine curve and the estimated sine curve in the estimation of the $\theta_M - \theta_X$ and $\theta_Y - \theta_M \theta_X$ relationships with the semiparametric approach ranged from 0 to 0.15. There was no significant difference among truncation levels. The truncation level at 5 was sufficient to capture the exponential relationship.
7. The polynomial nonlinear function did not recover the true exponential nonlinear relationship with the semiparametric approach. A significantly large difference was

detected between the estimated curve and the true curve. The large truncation level ($G = 200$) did not increase accuracy in recovering the true exponential curve.

8. The polynomial nonlinear function did not recover the true sine nonlinear relationship with the semiparametric approach. A significantly large difference was detected between the estimated curve and the true curve. The large truncation level ($G = 200$) did not increase accuracy in recovering the true sine curve.

This chapter included three sections. It began with a summary and discussion of the simulation study. Then, it provided a general conclusion and recommendations for applied researchers. Finally, it concluded by discussing the contributions and limitations of the current study.

Performance of the Parametric and Semiparametric SEM

Model convergence. Overall, the parametric and semiparametric SEMs achieved satisfactory convergence rates. The exponential function and sine function had higher convergence rates than the polynomial function in both the parametric and semiparametric approaches. This was reasonable given that the quadratic and interaction effects are difficult to converge. The results showed that more chains and longer iterations per chain help improve the convergence rate in the polynomial condition. Although the convergence rate was similar between the two approaches, the parametric approach took much less time to converge than the semiparametric approach because the mixture model posited in the semiparametric approach was more time-consuming in the Bayesian estimation.

Nonlinearity recoveries. Orthogonal polynomial was not used in the analysis. Similar polynomial nonlinearity recovery rates were detected between the parametric approach and the semiparametric approach with the large truncation level of 200. When truncation level was small

(5, 20), the parametric approach more accurately captured the polynomial nonlinearity. Nevertheless, similar results were not detected in the nonpolynomial nonlinearity. The semiparametric approach had a higher recovery rate than the parametric approach, and a small truncation level (5) was sufficient to capture the exponential and sine nonlinearity.

However, when the polynomial function was misspecified in the exponential and sine model, neither the parametric approach nor the semiparametric approach captured the true exponential and sine nonlinearity.

Conclusion and Recommendations

This article introduced the semiparametric Bayesian approach in estimating the direct effect of nonlinear functions in structural models with ordinal data. The performances of the parametric approach and the semiparametric approach were compared in the simulation study.

In conclusion, the semiparametric approach at truncation level 200 performed similarly to the parametric approach in recovering the polynomial nonlinear curves. However, the semiparametric approach at truncation level 200 was computationally heaved and time-consuming. Therefore, the parametric approach was suggested for application when a polynomial nonlinearity existed in the study.

In addition, the semiparametric approach had higher accuracy in capturing the nonpolynomial curves among latent factors in the structural model than the parametric approach. A lower truncation level (e.g., $G=5$) was sufficient to capture the nonlinearity. Thus, applied researchers were advised to use the semiparametric approach to detect the potential nonpolynomial relations among latent factors in the structural model.

However, when the nonlinear function was misspecified in the structural model, the semiparametric approach did not recover the true nonpolynomial relationship. A correctly

specified nonlinear function was strongly recommended for application in the model to accurately capture the nonlinearity.

Contributions and Limitations

In the current study, the performance of recovering the polynomial and nonpolynomial relationships was compared between the parametric approach and the semiparametric approach. This study extended the research field of current studies that limit the application of the semiparametric approach within the framework of nonnormality. To date, no study has had investigated whether the semiparametric approach accurately detects nonlinear relationships. The first contribution of this study was to explore whether the Bayesian semiparametric approach recovers polynomial and nonpolynomial nonlinearity in a latent structural model better than the parametric approach at different truncation levels varying from 5 to 200. Second, this study filled a gap in evaluating the recovery performance of the semiparametric approach when a polynomial function was misspecified in nonpolynomial data. It helped practitioners and researchers answer the practical question that significant quadratic and interaction terms do not warrant polynomial nonlinearity; instead, nonpolynomial nonlinearity could be a potential choice. Therefore, specifying a correct nonlinear function is critical in recovering the true nonlinearity. The results of this dissertation contributed to the literature as a reference for researchers and practitioners to select an appropriate truncation level when the semiparametric Bayesian approach is used to estimate different types of nonlinear relationships.

However, the current study also has several limitations. First, only three nonlinear functions were included in the study, which might not comprehensively represent all the nonlinear relationships among latent factors in the structural model. More nonlinear functions, such as the quartic nonlinear function, are expected to be investigated in future research. Second,

this study only tested whether the truncation level varies from 5 to 200. Shwaran and Zarepour (2000) found that as the truncation levels increase from 20 to 250, the truncation approximation of DP has much higher accuracy for detecting highly non-normal density. Therefore, a larger truncation level (e.g., 250) might lead to a better recovery rate in capturing polynomial nonlinearity as well as a better recovery of nonpolynomial nonlinearity when the nonlinear function is mis-specified. Third, the convergence of polynomial function was not very good. Future study may try to use orthogonal polynomial to stimulate the posterior convergence.

References

- Arminger, G., & Muthén, B. O. (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the metropolis-hastings algorithm. *Psychometrika*, *63*(3), 271-300. doi:10.1007/bf02294856
- Baron, R. M., & Kenny, D. A. (1986). The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173-1182. doi:10.1037/0022-3514.51.6.1173
- Bauer, D. J. (2005). A Semiparametric Approach to Modeling Nonlinear Relations Among Latent Variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(4), 513-535. doi:10.1207/s15328007sem1204_1
- Blozis, S. A. (2007). On Fitting Nonlinear Latent Curve Models to Multiple Variables Measured Longitudinally. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(2), 179-201. doi:10.1080/10705510709336743
- Bollen, K. A., & Paxton, P. (1998). Interactions of latent variables in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(3), 267-293. doi:10.1080/10705519809540105
- Cheung, G. W., & Lau, R. S. (2008). Testing Mediation and Suppression Effects of Latent Variables: Bootstrapping With Structural Equation Models. *Organizational Research Methods*, *11*(2), 296-325. doi:10.1177/1094428107300343
- Chow, S.-M., Tang, N., Yuan, Y., Song, X., & Zhu, H. (2011). Bayesian estimation of semiparametric nonlinear dynamic factor analysis models using the Dirichlet process prior. *British Journal of Mathematical and Statistical Psychology*, *64*(1), 69-106. doi:10.1348/000711010X497262

- Fahrmeir, L., & Raach, A. (2007). A Bayesian Semiparametric Latent Variable Model for Mixed Responses. *Psychometrika*, *72*(3), 327. doi:10.1007/s11336-007-9010-7
- Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, *1*(2), 209-230.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, Calif: Sage Publications.
- Harring, J. R., Weiss, B. A., & Hsu, J.-C. (2012). A comparison of methods for estimating quadratic effects in nonlinear structural equation models. *Psychological Methods*, *17*(2), 193-214. doi:10.1037/a0027539
- Hayes, A. F., & Preacher, K. J. (2010). Quantifying and Testing Indirect Effects in Simple Mediation Models When the Constituent Paths Are Nonlinear. *Multivariate Behavioral Research*, *45*(4), 627-660. doi:10.1080/00273171.2010.498290
- Hu, L.-t., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, *112*(2), 351-362. doi:10.1037/0033-2909.112.2.351
- Hutchison, D. (2018). Bayesian Psychometric Modelling R. Levy and R. Mislevy Boca Raton, Chapman and Hall–CRC. ISBN 978-1-439-88467-6. In (Vol. 181, pp. 550-550).
- Ishwaran, H., & James, L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, *96*(453), 161-173.
doi:10.1198/016214501750332758
- Ishwaran, H., & Zarepour, M. (2002). Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, *30*(2), 269-283. doi:10.2307/3315951
- Jaccard, J., & Wan, C. K. (1995). Measurement error in the analysis of interaction effects between continuous predictors using multiple regression: Multiple indicator and

- structural equation approaches. *Psychological Bulletin*, 117(2), 348-357.
doi:10.1037/0033-2909.117.2.348
- Kelava, A. & Brandt, H. (2009). Estimation of nonlinear latent structural equation models using the extended unconstrained approach. *Review of Psychology*, 16(2), 123-132.
- Kelava, A., & Brandt, H. (2014). A general non-linear multilevel structural equation mixture model. *Frontiers in Psychology*, 5(748). doi:10.3389/fpsyg.2014.00748
- Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, 96(1), 201-210. doi:10.1037/0033-2909.96.1.201
- Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, 65(4), 457-474. doi:10.1007/bf02296338
- Klein, A. G., & Muthén, B. O. (2007). Quasi-Maximum Likelihood Estimation of Structural Equation Models with Multiple Interaction and Quadratic Effects. *Multivariate Behavioral Research*, 42(4), 647-673. doi:10.1080/00273170701710205
- Kleinman, K. P., & Ibrahim, J. G. (1998). A Semiparametric Bayesian Approach to the Random Effects Model. *Biometrics*, 54(3), 921-938. doi:10.2307/2533846
- Lee, S. Y. (2007). Structural equation modeling: a Bayesian approach. In. Chichester, England Hoboken, NJ: Chichester, England Hoboken, NJ: John Wiley.
- Lee, & Song, Y. (2012). *Basic and advanced structural equation modeling: with applications in the medical and behavioral sciences*. Hoboken: Hoboken: Wiley.
- Levy, R., & Mislevy, R. J. (2016). *Bayesian psychometric modeling*. Chapman and Hall/CRC.
- MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7(1), 83-104. doi:10.1037/1082-989X.7.1.83

- Mallinckrodt, B., Abraham, W. T., Wei, M., & Russell, D. W. (2006). Advances in testing the statistical significance of mediation effects. *Journal of Counseling Psychology, 53*(3), 372-378. doi:10.1037/0022-0167.53.3.372
- Marsh, H. W., Wen, Z., & Hau, K.-T. (2004). Structural Equation Models of Latent Interactions: Evaluation of Alternative Estimation Strategies and Indicator Construction. *Psychological Methods, 9*(3), 275-300. doi:10.1037/1082-989X.9.3.275
- McDonald, R. P. (1967). Numerical methods for polynomial models in nonlinear factor analysis. *Psychometrika, 32*(1), 77-112. doi:10.1007/bf02289406
- Moses, L. E. (1952). Non-parametric statistics for psychological research. *Psychological Bulletin, 49*(2), 122-143. doi:10.1037/h0056813
- Ping, R. A. (1995). A Parsimonious Estimating Technique for Interaction and Quadratic Latent Variables. *Journal of Marketing Research, 32*(3), 336-347. doi:10.2307/3151985
- Plummer, M. (2003, March). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, No. 125.10).
- Qin, L. (2018). Estimating Nonlinear Indirect Effects in Bayesian Semiparametric Structural Equation Model. *Multivariate Behavioral Research, 53*(1), 130-131. doi:10.1080/00273171.2017.1404896
- R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA. URL <http://www.rstudio.com/>.
- Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of Semiparametric Regression

- Models for Repeated Outcomes in the Presence of Missing Data. *Journal of the American Statistical Association*, 90(429), 106-121. doi:10.1080/01621459.1995.10476493
- Ruppert, D., Wand, M. P., & Carroll, R. J. (2009). Semiparametric regression during 2003-2007. *Electronic journal of statistics*, 3, 1193.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34(4, Pt. 2), 100-100.
- Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica Sinica*, 4(2), 639-650.
- Sit, V., Poulin-Costello, M., & Bergerud, W. (1994). *Catalogue of curves for curve fitting* (p. 110). Forest Sciences Research Branch, Ministry of Forests.
- Song, X.-Y., Lu, Z.-H., Cai, J.-H., & Hak-Sing Ip, E. (2013). A Bayesian Modeling Approach for Generalized Semiparametric Structural Equation Models. *Psychometrika*, 78(4), 624-647. doi:10.1007/s11336-013-9323-7.
- Song, X.-Y., Pan, J.-H., Kwok, T., Vandemput, L., Ohlsson, C., & Leung, P.-C. (2010). A semiparametric Bayesian approach for structural equation models. *Biometrical Journal*, 52(3), 314-332. doi:10.1002/bimj.200900135
- Song, X.-Y., Xia, Y.-M., & Lee, S.-Y. (2009). Bayesian semiparametric analysis of structural equation models with mixed continuous and unordered categorical variables. *Statistics in Medicine*, 28(17), 2253-2276. doi:10.1002/sim.3612
- Song, X.-Y., Xia, Y.-M., Pan, J.-H., & Lee, S.-Y. (2011). Model Comparison of Bayesian Semiparametric and Parametric Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, 18(1), 55-72. doi:10.1080/10705511.2011.532720

- Stolzenberg, R. M. (1980). The Measurement and Decomposition of Causal Effects in Nonlinear and Nonadditive Models. *Sociological Methodology*, 11, 459-488. doi:10.2307/270872
- Su, Y. S., & Yajima, M. (2012). R2jags: a package for running JAGS from R. R package version 0.03-08. <http://cran.r-project.org/package=R2jags>.
- West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In *Structural equation modeling: Concepts, issues, and applications*. (pp. 56-75). Thousand Oaks, CA, US: Sage Publications, Inc.
- Xia, Y., & Gou, J. (2016). Bayesian semiparametric analysis for latent variable models with mixed continuous and ordinal outcomes. *Journal of the Korean Statistical Society*, 45(3), 451-465. <https://doi.org/10.1016/j.jkss.2016.01.005>
- Yang, M., Dunson, D. B., & Baird, D. (2010). Semiparametric Bayes hierarchical models with mean and variance constraints. *Computational Statistics & Data Analysis*, 54(9), 2172-2186. <https://doi.org/10.1016/j.csda.2010.03.025>
- Zeger, S. L., & Karim, M. R. (1991). Generalized Linear Models with Random Effects; a Gibbs Sampling Approach. *Journal of the American Statistical Association*, 86(413), 79-86. doi:10.1080/01621459.1991.10475006