A CLASS OF DISTRIBUTION-FREE MODELS FOR LONGITUDINAL MEDIATION ANALYSIS

D. Gunzler,
CASE WESTERN RESERVE UNIVERSITY AT METROHEALTH MEDICAL CENTER

W. Tang,
UNIVERSITY OF ROCHESTER

N. Lu,
UNIVERSITY OF ROCHESTER AND CANANDAIGUA VA MEDICAL CENTER

P. Wu, and
CHRISTIANA CARE HEALTH SYSTEM

X.M. Tu
UNIVERSITY OF ROCHESTER AND CANANDAIGUA VA MEDICAL CENTER

Abstract

Mediation analysis constitutes an important part of treatment study to identify the mechanisms by which an intervention achieves its effect. Structural equation model (SEM) is a popular framework for modeling such causal relationship. However, current methods impose various restrictions on the study designs and data distributions, limiting the utility of the information they provide in real study applications. In particular, in longitudinal studies missing data is commonly addressed under the assumption of missing at random (MAR), where current methods are unable to handle such missing data if parametric assumptions are violated.

In this paper, we propose a new, robust approach to address the limitations of current SEM within the context of longitudinal mediation analysis by utilizing a class of functional response models (FRM). Being distribution-free, the FRM-based approach does not impose any parametric assumption on data distributions. In addition, by extending the inverse probability weighted (IPW) estimates to the current context, the FRM-based SEM provides valid inference for longitudinal mediation analysis under the two most popular missing data mechanisms; missing completely at random (MCAR) and missing at random (MAR). We illustrate the approach with both real and simulated data.

Keywords

mediation analysis; functional response models; structural equation model; weighted generalized estimating equations II; missing data; generalized least squares

1. Introduction

Mediational analysis is employed in a wide range of behavioral, biomedical, and psychosocial research studies to investigate the causal mechanism of intervention, i.e., mediation. The traditional regression paradigm is ill-suited for modeling such multidimensional causal relationships, because of the complex relationships among different variables and multiple roles a variable may play. Although the structural equation model (SEM) provides an ideal conceptual framework for the dynamic role a mediator plays in channeling the effect of a causal agent to the outcome of interest, the various restrictions imposed by the available inference methods have hindered its wide use and limited the utility of the information provided in real studies.

In real longitudinal studies, we often encounter missing data on the mediator and outcome of interest that is dependent on information from a previous time point, an assumption that has been termed missing at random (MAR). For instance, if an intervention is effective, a patient may drop out of a study if they believe there is no added benefit in continuing the treatment. The missing data is associated with the treatment. If parametric assumptions are not met, under this commonly used assumption of MAR, there currently is no method available for unbiased inference in the SEM framework in the context of mediation analysis.

In this paper, we discuss a new, robust approach to address the various limitations of existing methods by utilizing a class of functional response models (FRM). In Section 2, we briefly review the concept of mediation and the SEM modeling framework for such a causal process. In Section 3, we discuss how to frame the mediation model using FRM after giving a brief review of the latter. In Section 4, we discuss inference for the distribution-free FRM-based SEM for both complete and missing data in the longitudinal study design. In Section 5, we illustrate the proposed approach with both real and simulated data, and compare its performance with existing alternatives. In Section 6, we give our concluding remarks.

2. Structural Equations Model for Mediation Analysis

2.1. Mediation Analysis

In treatment studies, it is often of great interest to identify and study mechanisms by which an intervention achieves its effect. By investigating such a mediational process through which the treatment affects study outcomes, not only can we further our understanding of the pathology of the disease and treatment, but may also provide alternative intervention strategies for the disease with efficient use of resources. For example, a tobacco prevention program may teach participants how to stop taking smoking breaks at work, thereby changing the social norms for tobacco use. As a result, this change in social norms reduces cigarette smoking (MacKinnon & Fairchild, 2009). With mediation analysis, we gain insight and acquire deep understanding about the mechanism of action of pharmacological and psychotherapeutic treatments. Such information provides an added dimension to understand the etiology of disease and pathways of therapeutic effects so more efficacious and cost efficient alternative therapies may be developed.

Baron and Kenny (1986), in the first paper addressing mediation analysis, tested the mediation process through a series of regression equations. Since variables in a causal relationship can play both roles of the cause and effect, the standard regression paradigm is ill-suited for modeling such a relationship, because of the unique cause or effect qualifier of each variable it mandates (Kraemer, 2001; MacKinnon & Fairchild, 2009; Rothman & Greenland, 1998). The structural equation model (SEM) provides a formal modeling and inference framework for mediation and other related causal analysis.

There are many advantages for using the SEM framework in the context of mediation analysis. SEM simplifies testing of mediation hypotheses as SEM is designed, in part, to test these more complicated mediation models in a single analysis (MacKinnon, 2008). In standard regression, ad hoc methods such as the ones developed by Baron and Kenny (1986), Sobel (1982), and Clogg, Petkova, and Shihadeh (1992) must be used for inference about indirect and total effects. These ad hoc methods rely on combining the results of two or more equations to derive the asymptotic variance. This is especially problematic under missing data where a different number of observations may be missing in the different regression equations representing a mediation process. Also, in standard regression, by default, we handle missing data via listwise deletion since there is no built in missing data mechanism when using ordinary least squares (OLS).

Further, since we make causal assumptions in a mediation model, SEM analysis approach provides model fit information about consistency of the hypothesized mediational model to the data and evidence for the plausibility of the causality assumptions (Bollen & Pearl, 2012; Imai, Keele, & Tingley, 2010) we make in a mediation model. The procedure by Baron and Kenny (1986) has also been shown to be low powered (MacKinnon, 2008). SEM allows for ease of extension to longitudinal data within a single framework, corresponding with a study's conceptual framework for clear hypothesis articulation (Preacher, Wichman, MacCallum, & Briggs, 2008). Finally, Bollen and Pearl (2012) note that even the same equation in SEM and regression analysis is based on completely different assumptions and has a different meaning. Standard regression analysis implies a statistical relationship based on a conditional expected value, while SEM implies a functional relationship expressed via a conceptual model, path diagram, and mathematical equations. The causal relationships in a hypothesized mediation process, simultaneous nature of the indirect and direct effects and the dual role the mediator plays as both a cause for the outcome and effect of the intervention can therefore be expressed better in terms of structural equations than regression analysis.

2.2. Structural Equations Model (SEM)

The structural equation model (SEM) addresses complex relationships among variables that are generally depicted in path diagrams. A path diagram consists of nodes representing the different variables and arrows showing relations among the variables. For example, shown in Figure 1 is the path diagram for the causal relationship between the three variables in the smoking prevention example discussed earlier. The three variables, prevention intervention (x_{i1}), social norm (z_{i2}), and amount of smoking (y_{i3}), are measured at three assessment points in chronological order starting at baseline within a longitudinal setting. Also, all

effect-receiving variables such as the social normal and amount of smoking in this example are called endogenous variables, and effect-imparting variables such as the prevention intervention are known as exogenous variables in the nomenclature of SEM. In this paper, all variables are observed unless otherwise noted. See Bollen (1989) and Kowalski and Tu (2007) for more details about modeling complex relationships involving latent constructs using SEM.

Consider a longitudinal study with three assessment times indexed by t $(1 \leq t \leq 3)$. Let x_{it}, z_{it}, and y_{it} denote the causal (or predictor), mediator and response variables at time t. In this paper, we focus on continuous mediators and responses. The SEM for a typical mediational process involving a single predictor, mediator and response as depicted in the path diagram in Figure 1 is given by

$$
z_{i2}=\beta_0+\beta_{xz}x_{i1}+\epsilon_{zi}, \qquad\qquad y_{i3}=\gamma_0+\gamma_{xy}x_{i1}+\gamma_{zy}z_{i2}+\epsilon_{yi},
$$
$$
\epsilon_i=\begin{pmatrix}\epsilon_{yi}\\ \epsilon_{zi}\end{pmatrix}\sim(\mathbf{0},\mathbf{\Psi}),\quad \mathbf{\Psi}=\begin{pmatrix}\sigma^2_{\epsilon_y} & \sigma_{\epsilon_y\epsilon_z}\\ & \sigma^2_{\epsilon_z}\end{pmatrix}, \qquad x_{i1}\perp\epsilon_{zi},\qquad x_{i1},z_{i2}\perp\epsilon_{yi}. \quad\text{(1)}
$$

The error terms ϵ_{yi} and ϵ_{zi} are typically assumed to be normal. However, since we focus on robust inference, we only assume that ϵ_{yi} and ϵ_{zi} are both continuous with mean 0 and variance $\mathbf{\Psi}$. The above may be expressed in general matrix notation as follows:

$$
\mathbf{u}_i=B\mathbf{u}_i+\Gamma\mathbf{x}_i+\epsilon_i,\quad \epsilon_i\sim(\mathbf{0},\mathbf{\Psi}),\mathbf{x}_i\perp\epsilon_i, \quad\text{(2)}
$$

or alternatively,

$$
\mathbf{u}_i=(\mathbf{I}_2-B)^{-1}\Gamma\mathbf{x}_i+\widetilde{\epsilon}_i,\qquad \widetilde{\epsilon}_i\sim(\mathbf{0},\mathbf{\Phi}),\mathbf{x}_i\perp\widetilde{\epsilon}_i, \quad\text{(3)}
$$

where

$$
\mathbf{u}_i=(y_{i3},z_{i2})^\top, \qquad \mathbf{x}_i=(1,x_{i1})^\top, \qquad \theta=\left(\gamma_0,\gamma_{zy},\gamma_{xy},\beta_0,\beta_{xz},\sigma^2_{\epsilon_z},\sigma_{\epsilon_y\epsilon_z},\sigma^2_{\epsilon_y}\right)^\top,
$$
$$
\mathbf{I}_2=\begin{pmatrix}1&0\\0&1\end{pmatrix},\qquad B=\begin{pmatrix}0&\gamma_{zy}\\0&0\end{pmatrix},\qquad \Gamma=\begin{pmatrix}\gamma_0&\gamma_{xy}\\\beta_0&\beta_{xz}\end{pmatrix}, \qquad\qquad\text{(4)}
$$
$$
\widetilde{\epsilon}_i=(\mathbf{I}_2-B)^{-1}\epsilon_i, \qquad\qquad \mathbf{\Phi}=(\mathbf{I}_2-B)^{-1}\mathbf{\Psi}(\mathbf{I}_2-B)^{-\top}.
$$

Note that stochastic independence is not taken for granted, as it is particularly important for causal inference (Bollen & Pearl, 2012; Imai et al., 2010). To facilitate validation, the usual independence is replaced by zero correlation, or *pseudo-isolation*, which can be empirically checked. For example, to assess the causal effect of x_{i1} and z_{i2} in (1), it is critical that x_{i1} be uncorrelated with ϵ_{zi} in the first and both x_{i1} and z_{i2} be uncorrelated with ϵ_{yi} in the second equation of the SEM. It is then readily checked that

$$
Cov\left(\epsilon_{zi},\epsilon_{yi}\right)=Cov\left(\epsilon_{yi},z_{i2}\right)=0. \quad\text{(5)}
$$

In other words, ϵ_{zi} and ϵ_{yi} are uncorrelated as well, i.e., $\sigma_{\epsilon_y\epsilon_z}=0$, for this particular SEM, and thus θ in (4) reduces to $\theta=\left(\gamma_0,\gamma_{zy},\gamma_{xy},\beta_0,\beta_{xz},\sigma^2_{\epsilon_z},\sigma^2_{\epsilon_y}\right)^\top$.

Note that SEM such as the one in (1) only provides causal inference in the absence of selection bias, such as data from randomized controlled trials. Imai et al. (2010) have proposed approaches to extend SEM in the absence of selection bias based on the counterfactual outcome based causal framework, such as data from most epidemiological studies. In this paper, we assume no selection bias unless otherwise stated.

Although (1) can also be applied to cross-sectional data, the longitudinal data setting is more popular for mediation analysis since temporal changes are important for modeling causal relationships. In this paper we consider longitudinal data from multiple waves, based on a hypothesized relationship between x_{i1}, z_{i2}, and y_{i3}. These repeated measures from different times are linked not only with each other, but with other variables as well (see Section 5.1.1). For example, under MAR the missingness of z_{i2} may depend on z_{i1} and even y_{i1} in addition to x_{i1} (Kowalski & Tu, 2007; Robins, Rotnitzky, & Zhao, 1995).

For mediation analysis, the primary interest is the hypothesis of full mediation $H_0 : \gamma_{xy} = 0$. Under this null, the direct path from x to y is broken, with the effect of x on y fully mediated through the change in z. In practice, however, it is more common that a researcher comes across partial mediation, where the direct path from x to y is partially broken through the change in z, indicating that z mediates some of the effect of x on y. Under partial mediation, one becomes interested in constructing the direct, γ_{xy}, indirect, $\beta_{xz}\gamma_{xy}$, and total, $\gamma_{xy} + \beta_{xz}\gamma_{xy}$, effect of the predictor x_{i1}. The three types of effects are readily obtained from estimates of θ. Inference (standard errors and p-values) about such effects is also easily performed using the Delta or Bootstrap methods (e.g., Sobel, 1982; Clogg et al., 1992; Bollen & Stine, 1990).

Significant advances have been made over the past few decades in the theory and applications as well as software development for fitting SEM models. For example, in addition to specialized packages such as LISREL (Joreskog & Sorbom, 1997), MPlus (Muthén & Muthén, 1998–2010), EQS (Bentler, 2006), and Amos (Arbuckle, 1995–2010), procedures for fitting SEM are also available from general-purposes statistical packages such as R, SAS, STATA, and Statistica. These packages provide inference based on maximum likelihood (ML), generalized least squares (GLS), and weighted least squares (WLS). While ML assumes a multivariate normal for the joint distribution of all variables (exogenous plus endogenous variables such as the trivariate $(x_{i1}, z_{i2}, y_{i3})^\top$ in (1) within our context), GLS and WLS do not, thereby providing more robust estimates.

In recent years, many software packages have implemented robust variance estimates to improve the validity of inference in the presence of departures from assumed parametric models (Maas & Hox, 2004; Van der Leeden & Busing, 1994; Van der Leeden, Busing, & Meijer, 1997). In most real studies, missing data invariably occurs, especially for longitudinal studies. In the presence of missing data, such variance estimates not only do little to prevent inference bias (wrong standard errors and type I error levels), but also give rise to bias in parameter estimates (Lu, Tang, He, Yu, Crits-Christoph, Zhang, & Tu, 2009). Our simulations for the SEM-based mediation model in (1) have demonstrated similar poor performances by the GLS, ML, and WLS under departures from the assumed normality in the presence of missing data. We discuss these findings in detail in Section 5.

Other approaches have also been proposed to address missing data. For example, Papadopoulos and Amemiya (2005) proposed a multistage analysis approach to handle two nonnormal, correlated endogenous variables and unbalanced data (due to missing data) in the longitudinal setting (Allison, 1987; Baker & Fulker, 1983; Satorra, 2002; Werts, Rock, & Grandy, 1979). However, their approach to missing data is analogous to the generalized estimating equations (GEE), which provides valid inference only for a more limited class of missing data following the missing completely at random (MCAR), rather than the more general response-dependent missing at random (MAR) mechanism focused in this paper.

In a recent paper, Zhang and Wang (2013) introduced a method using normal data and the EM algorithm along with bootstrapping for mediation analysis. However, like other methods that rely on parametric assumptions, their approach will also yield invalid inference under MAR, if the parametric assumptions are violated.

Finally, outside of the SEM framework using the Baron and Kenny (1986) approach, we might observe similar point estimates, though slightly different standard errors and Type I error, via OLS and SEM ML under complete data or under the MCAR assumption when standard regression assumptions are met. However, OLS does not have a built in mechanism to handle missing data and does not provide robust inference under the MAR assumption.

However, unlike traditional maximum likelihood inference for regression models, where only the regression coefficients such as γ_0, γ_{xy}, γ_{zy}, β_0, and β_{xz} are estimated, the ML for SEM must also estimate the variance parameters such as $\sigma^2_{\epsilon_y}$ and $\sigma^2_{\epsilon_z}$ in (1) in order to make the γ's identifiable (Bollen, 1989; Gunzler, 2011). However, the necessity to estimate these parameters not only complicates the estimation procedure, but also renders conventional robust methods inapplicable to SEM.

Before introducing a distribution-free approach to address the aforementioned flaws of available methods within a longitudinal data setting, we first give a brief overview of the functional response models upon which the new approach is premised.

3. Functional Response Models

As we have discussed earlier, we use SEM methods for mediation analysis in this paper that make simultaneous inference about both equations in the mediation model. We develop an approach in this section that will be used to achieve the inference step within the context of distribution-free SEM. Traditional regression models for cross-sectional data will not provide such inference and are all defined based on a single-subject response. For example, consider a sample of size n, and let y_i and $\mathbf{x}_i$ denote some response and a vector of explanatory variables (independent variables, predictors, or covariates) of interest ($1 \quad i \quad n$). The popular linear model is defined by, $E\left(y_i|\mathbf{x}_i\right)=\mathbf{x}_i^\mathsf{T}\beta$, where $E(y_i \mid \mathbf{x}_i)$ denotes the conditional mean of y_i given $\mathbf{x}_i$, and β a vector of parameters. In this model, the dependent variable is a single-subject response y_i. Although the linear model has been extended for modeling more complex types of response variables such as binary and count data (McCullagh & Nelder, 1989), such models still involve the single-subject response y_i. For example, in the generalized linear model defined by $E\left(y_i|\mathbf{x}_i\right)=g\left(\mathbf{x}_i^\mathsf{T}\beta\right)$, although the right

side is generalized to be a function of the linear predictor, $\mathbf{x}_i^\top \beta$, to accommodate the range restriction of non-linear response y_i, the left side remains identical to the linear model. Further, the response only appears on the left, while the explanatory variable must stay on the right side of the model.

These constraints prevent regression models from being applied to many problems of interest in practice. For example, within our context of SEM, since an endogenous variable can be on either side of an equation, this feature immediately excludes regression as a framework for inference about SEM. In other applications such as correlation analysis, mixture models and social network connectivity, we are interested in relationships between defined by outcomes from multiple subjects, which again violates the confines of regression. By addressing the fundamental limitations of the classic regression, the *functional response model* (FRM) can express a broader class of problems under a regression-like framework (Kowalski & Tu, 2007; Kowalski, Pagano, & DeBruttola, 2002; Kowalski & Powell, 2004; Ma, Tang, Feng, & Tu, 2008; Ma, Tang, Yu, & Tu, 2010; Tu, Feng, Kowalski, Tang, Wang, Wan, & Ma, 2007; Yu, Tang, Kowalski, & Tu, 2011; Yu, Chen, Tang, He, Gallop, Crits-Christoph, Hu, & Tu, 2013).

Consider a class of distribution-free regression models defined by

$$
\begin{aligned}
E \quad & \left[\mathbf{f} \left(\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_q} \right) \mid \mathbf{g} \left(\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_q} \right), \mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_q} \right] \\
=& \mathbf{h} \left(\mathbf{g} \left(\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_q} \right), \mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_q}; \theta \right), \\
& \left(i_1, \ldots, i_q \right) \in C_q^n, 1 \leq q, \; 1 \leq i \leq n,
\end{aligned}
\tag{6}
$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{iq})^\top$ denotes the vector of responses from the ith subject, $\mathbf{f}$ is some vector-valued function and $\mathbf{g}$ is a subset of variables $\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_q}$ with $\mathbf{f}$ and $\mathbf{g}$ containing nonoverlapping variables, $\mathbf{h}(\theta)$ some vector-valued smooth function (e.g., with continuous derivatives up to the second order), θ a vector of parameters of interest, q some positive integer, and C_q^n the set of $\begin{pmatrix} n \\ q \end{pmatrix}$ combinations of q distinct elements $(i_1, \ldots, i_q)$ from the integer set $\{1, \ldots, n\}$. FRM in (6) extend the single-subject response in the classic GLM to a function of responses from multiple subjects. For example, by setting $q = 1$, we immediately obtain from (6) the class of distribution-free GLM.

To apply the FRM in our setting, let

$$
\begin{aligned}
\mathbf{f}_i \quad &= (\mathbf{f}_{1i}^\top, \mathbf{f}_{2i}^\top)^\top, \qquad \mathbf{h}_i (\theta) = \mathbf{h} (\mathbf{x}_i, \theta) = (\mathbf{h}_{1i}^\top (\theta), \mathbf{h}_{2i}^\top (\theta))^\top, \\
& i = 1, 2, \ldots, n, \\
\mathbf{f}_{1i} &= (y_{i3}, z_{i2})^\top, \qquad \mathbf{f}_{2i} = (y_{i3}^2, y_{i3} z_{i2}, z_{i2}^2)^\top, \qquad \mathbf{x}_i = x_{i1}, \\
\mathbf{h}_{1i} (\theta) &= ((\gamma_0 + \gamma_{zy} \beta_0) + (\gamma_{xy} + \gamma_{zy} \beta_{xz}) x_{i1}, \; \beta_0 + \beta_{xz} x_{i1})^\top, \\
\mathbf{h}_{2i} (\theta) &= E (\mathbf{f}_{2i} \mid \mathbf{x}_i) = (E (y_{i3}^2 \mid \mathbf{x}_i), E (y_{i3} z_{i2} \mid \mathbf{x}_i), E (z_{i2}^2 \mid \mathbf{x}_i))^\top,
\end{aligned}
\tag{7}
$$

where $\theta = \left(\gamma_0, \gamma_{zy}, \gamma_{xy}, \beta_0, \beta_{xz}, \sigma_{\epsilon_z}^2, \sigma_{\epsilon_y}^2 \right)^\top$ is the set of parameters, and

$$
\begin{aligned}
E\left(z_{i2}^2 \mid \mathbf{x}_i\right) &= \sigma_{\epsilon_z}^2 + \left(\beta_0 + \beta_{xz} x_{i1}\right)^2, \\
E\left(y_{i3} z_{i2} \mid \mathbf{x}_i\right) &= \gamma_{zy}\left(\beta_0 + \beta_{xz} x_{i1}\right)\left(\beta_0 + \beta_{xz} x_{i1}\right) + \gamma_{zy}\sigma_{\epsilon_z}^2 \\
&\quad + \left(\beta_0 + \beta_{xz} x_{i1}\right)\left[\left(\gamma_0 + \gamma_{zy}\beta_0\right) + \left(\gamma_{xy} + \gamma_{zy}\beta_{xz}\right) x_{i1}\right], \\
E\left(y_{i3}^2 \mid \mathbf{x}_i\right) &= \gamma_{zy}^2 \sigma_{\epsilon_z}^2 + \sigma_{\epsilon_y}^2 + \left[\left(\gamma_0 + \gamma_{zy}\beta_0\right) + \left(\gamma_{xy} + \gamma_{zy}\beta_{xz}\right) x_{i1}\right]^2.
\end{aligned}
\tag{8}
$$

Then the FRM for the SEM in (7) is

$$
E\left(\mathbf{f}_i \mid \mathbf{x}_i\right) = \mathbf{h}_i\left(\theta\right), \quad i = 1, 2, \ldots, n. \tag{9}
$$

Note that as $\sigma_{\epsilon_y \epsilon_z} = 0$ following from (1), this parameter is not included in θ. Note also that the $\mathbf{x}_i$ in the FRM above is defined differently from the $\mathbf{x}_i$ in (2) or (3).

As in the case of (1), $\mathbf{h}_i(\theta)$ above is not linear in the components of θ, and thus the FRM-based SEM in (9) is again not a linear model. The first component $\mathbf{f}_{1i}$ of the response function $\mathbf{f}_i$ is identical to $\mathbf{u}_i$ in the parametric setup in (1), while the second $\mathbf{f}_{2i}$ contains the necessary second-order moments of $\mathbf{u}_i$ to provide the needed information to identify all the γ's in (7). The three higher-order moment equations in (8) not only permit estimation of the two variance parameters $\sigma_{\epsilon_y}^2$ and $\sigma_{\epsilon_z}^2$, but also supplement the needed information to address the identifiability of the γ's.

For the particular SEM in (1) for mediation analysis, we may also define an alternative FRM to estimate the parameters of primary interest, $\theta = (\gamma_0, \gamma_{xy}, \gamma_{zy}, \beta_0, \beta_{xz})^\top$, without the help of higher-order moments as $\mathbf{f}_{2i}$ in (7). The issue of identifiability is introduced when taking the conditional expectation of y_{i3} given x_{i1} in the second equation (Bollen, 1989; Gunzler, 2011). This "smoothing" process eliminates the independent information needed to estimate γ_{zy}, thereby rendering the resulting smoothed version incapable of identifying the parameters in the second regression model in (1). Thus, to identify all the γ's without relying on higher-order moments, we may define an FRM by bypassing such a smoothing process.

To this end, consider an FRM defined by

$$
\begin{aligned}
E\quad & \left(f_{i1} \mid \mathbf{g}_{i1}\right) = h_1\left(\mathbf{g}_{i1}, \theta\right), \qquad E\left(f_{i2} \mid \mathbf{g}_{i2}\right) = h_2\left(\mathbf{g}_{i2}, \theta\right), \\
& f_{i1} = y_{i3}, \; f_{i2} = z_{i2}, \; \mathbf{g}_{i1} = \{x_{i1}, z_{i2}\}, \; \mathbf{g}_{i2} = x_{i1}, \\
& h_1\left(\mathbf{g}_{i1}, \theta\right) = \gamma_0 + \gamma_{xy} x_{i1} + \gamma_{zy} z_{i2}, \; h_2\left(\mathbf{g}_{i2}, \theta\right) = \beta_0 + \beta_{xz} x_{i1}.
\end{aligned}
\tag{10}
$$

As the SEM in (9), f_{i1} and $\mathbf{g}_{i1}$ contain no common variable. Since the mean response $h_{i1}(\theta)$ above retains the variable z_{i2}, all model parameters θ are uniquely defined, and thus estimable. Further, the FRM is not a longitudinal regression, since $\mathbf{g}_{it}$ involves a varying set of variables for $t = 1, 2$.

For either FRM-based mediation model in (7) and (10), its cross-sectional data version is similarly defined by simply modifying the respective components of the FRMs. Since the alternative FRM in (10) does not involve second-order moments as does ML, GLS, and WLS, it is much easier to implement, saving valuable computation time.

4. Distribution-Free Inference

We first discuss distribution-free inference with no missing data, which is the same for either cross-sectional or longitudinal data, and then extend the considerations to the longitudinal setting under the two most common missing data mechanisms, MCAR and MAR.

4.1. Complete Data

Consider the FRM-based mediation model in (7) with complete data, with $\mathbf{f}_i = \left(y_{i3}, z_{i2}, y_{i3}^2, y_{i3}z_{i2}, z_{i2}^2 \right)^{\mathsf{T}}$. Let

$$S_i = \mathbf{f}_i - \mathbf{h}_i, \qquad D_i = \frac{\partial}{\partial\theta}\mathbf{h}_i, \qquad V_i = Var\left(\mathbf{f}_i \mid \mathbf{x}_i \right), \qquad \mathbf{w}_{ni} = D_i V_i^{-1} S_i. \quad (11)$$

We estimate θ by the following set of estimating equations:

$$\mathbf{w}_n\left(\theta\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_{ni} = \sum_{i=1}^{n} D_i V_i^{-1} S_i = \mathbf{0}. \quad (12)$$

Given the model specifications, S_i and D_i are readily evaluated. Further, given parametric assumptions in (1), V_i is also easily computed. Under (9), the estimate $\hat{\theta}$ obtained as the solution of θ to the equations in (12) is consistent and asymptotically normal, regardless of the data distributions, i.e.,

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \rightarrow_d N\left(0, \Sigma_\theta\right), \quad B = E\left(D_i V_i^{-1} D_i^{\mathsf{T}} \right), \quad \Sigma_\theta = B^{-1} E\left(D_i V_i^{-1} S_i S_i^{\mathsf{T}} V_i^{-1} D_i^{\mathsf{T}} \right) B^{-\mathsf{T}}, \quad (13)$$

where $\rightarrow_d$ denotes convergence in distribution.

Unlike the ML estimate, the asymptotic results above do not require the (normal) distribution assumptions. But if the conditional joint $\mathbf{u}_i = (y_{i3}, z_{i2})^{\mathsf{T}}$ given $\mathbf{x}_{i1}$ does follow the multivariate normal, the asymptotic variance Σ_θ in (13) simplifies to the *model-based* asymptotic variance $\Sigma_\theta = B^{-1}$. A consistent estimate of Σ_θ is obtained by substituting moment estimates in place of the respective parameters:

$$\hat{\Sigma}_\theta = \hat{B}^{-1}\left(\frac{1}{n-1}\sum_{i=1}^{n} \hat{D}_i \hat{V}_i^{-1} \hat{S}_i \hat{S}_i^{\mathsf{T}} \hat{V}_i^{-1} \hat{D}_i^{\mathsf{T}} \right) \hat{B}^{-\mathsf{T}}, \quad \hat{B} = \frac{1}{n-1}\sum_{i=1}^{n} \hat{D}_i \hat{V}_i^{-1} \hat{D}_i^{\mathsf{T}},$$

where $\hat{B}_i$, $\hat{D}_i$, $\hat{S}_i$, and $\hat{V}_i$ denote the corresponding quantities with θ replaced by $\hat{\theta}$. Further, since we are fixing on our exogenous variable x_{i1} in our estimating equations, the asymptotic results of the FRM-based approach will hold regardless of what type of variable (i.e., discrete, continuous) or the data distribution of x_{i1}.

The set of equations in (12) bears a close resemblance to the generalized estimating equations II or GEE II (Kowalski & Tu, 2007; Prentice & Zhao, 1991; Reboussin & Liang, 1998). For convenience, we refer to (12) as GEE II throughout the rest of discussion. By setting $\mathbf{f}_i$ and $\mathbf{h}_i$ equal to the ones in (10), we can also use (12) to provide inference about θ

for this alternative FRM. To ensure consistent estimates, however, we must select V_i in such a way that $E\left(V_i^{-1}S_i|\mathbf{x}_{i1}\right)=V_i^{-1}E\left(S_i+\mathbf{x}_{i1}\right)$ (Scharfstein, Rotnitzky, & Robins, 1999). For example, this condition is met, if $V_i=\text{Var}(\mathbf{f}_i\,|\,x_{i1})$. But, if $V_i=\text{Var}(\mathbf{f}_i\,|\,x_{i1},\,z_{i2})$, $E\left(V_i^{-1}S_i|\mathbf{x}_{i1}\right)\neq V_i^{-1}E\left(S_i|\mathbf{x}_{i1}\right)$ and (12) is not guaranteed to provide consistent estimates. Another choice of V_i to ensure this condition is

$$V_i=\begin{pmatrix} Var\left(f_{i1}\,|\,x_{i1}\right) & Cov\left(f_{i1},f_{i2}\,|\,x_{i1}\right) \\ Cov\left(f_{i1},f_{i2}\,|\,x_{i1}\right) & Var\left(f_{i2}\,|\,x_{i1},z_{i2}\right) \end{pmatrix}.$$

A straightforward choice of V_i to ensure this condition is $\mathbf{I}_2$. The estimating equations approach discussed yields more robust inference than ML. Therefore, using nonnormally distributed data, we will obtain more robust estimates for the standard errors and Type I errors using this approach. Also of note is that while we are only estimating first moment parameters in the alternative FRM, the estimation procedure accounts for the correlated responses because we are including a working covariance matrix V_i.

4.2. Missing Data

Applications of GEE II are also used for unbalanced data, and analogous methods to GEE II have been proposed for nonnormal, correlated, and unbalanced data in the longitudinal setting, applicable to mediation analysis (Allison, 1987; Baker & Fulker, 1983; Papadopoulos & Amemiya, 2005; Satorra, 2002; Werts et al., 1979). However, as missing data with an underlying mechanism for missingness arise in virtually all real studies, applications of GEE II when restricted to the subsample of subjects with completely observed data over all assessments ($t = 1, 2, 3$) are not only inefficient, but more importantly are also vulnerable to selection bias (Kowalski & Tu, 2007; Lu et al., 2009; Robins et al., 1995; Scharfstein et al., 1999; Tsiatis, 2006).

For mean-based distribution-free models such as the generalized linear model, the weighted generalized estimating equations (WGEE) is the most popular for inference when the missing data follows the missing at random (MAR) assumption, a plausible and yet quite general statistical model for missing data that is applicable to many studies in practice (Kowalski & Tu, 2007; Little & Rubin, 1987; Lu et al., 2009; Robins et al., 1995; Scharfstein et al., 1999; Tsiatis, 2006). We discuss below how to extend this inverse probability weighting (IPW) based approach to the current FRM-based SEM models for mediation analysis.

We assume that there is no missing data at baseline $t = 1$, and missing data is the result of missed visit. Thus, the trio (x_{it}, z_{it}, y_{it}) is observed for $t = 1$, but subject to missing for $t = 2$, 3. Define a missing (or rather observed) data indicator for each ith subject as follows:

$$r_{it}=\begin{cases} 1 & \text{if}\,(x_{it},z_{it},y_{it})\,\text{is observed} \\ 0 & \text{if}\,(x_{it},z_{it},y_{it})\,\text{is missing} \end{cases}, \quad \mathbf{r}_i=(r_{i1},r_{i2},r_{i3})^{\intercal}.$$

Under the assumptions, $r_{i1} = 1$ ($1 \le i \le n$). Let

$$\pi_{it} = Pr\left(r_{it}=1 \mid \mathbf{x}_i, \mathbf{z}_i, \mathbf{y}_i\right), \quad \Delta_{it} = \frac{r_{it}}{\pi_{it}}. \quad (14)$$

In most applications, the missing data probability π_{it} is unknown and must be estimated. Under the missing completely at random (MCAR) assumption (Little & Rubin, 1987), $\mathbf{r}_i$ is independent of $\mathbf{x}_i$, $\mathbf{z}_i$ and $\mathbf{y}_i$, yielding $\pi_{it} = \Pr(r_{it} = 1) = \pi_t$ ($2 \le t \le 3, 1 \le i \le n$). In this case, π_t, a constant independent of $\mathbf{x}_i$, $\mathbf{z}_i$ and $\mathbf{y}_i$, is readily estimated by the sample moment:

$$\hat{\pi}_t = \frac{1}{n}\sum_{i=1}^{n} r_{it} \ (2 \le t \le 3).$$

Under MAR, π_{it} becomes dependent on the observed x_{is}, z_{is}, y_{is} for s up to and including time t ($2 \le t \le 3$), making it difficult to model and estimate π_{it} without imposing some further assumption. One popular constraint is the Monotone Missing Data Pattern (MMDP). Under MMDP, x_{it}, z_{it} and y_{it} are observed only if their predecessors x_{is}, z_{is}, and y_{is} prior to time t are all observed. As patient dropout is the most popular MAR, this assumption not only reduces the number of missing data patterns and complexity in modeling π_{it}, but also posits a plausible model for modeling missingness in most real studies (Kowalski & Tu, 2007; Robins et al., 1995).

For $2 \le t \le 3$, let

$$\begin{aligned}
\widetilde{\mathbf{x}}_{it} &= \left(x_{i1},\ldots,x_{i(t-1)}\right)^{\mathsf{T}}, & \widetilde{\mathbf{z}}_{it} &= \left(z_{i1},\ldots,z_{i(t-1)}\right)^{\mathsf{T}}, \\
\widetilde{\mathbf{y}}_{it} &= \left(y_{i1},\ldots,y_{i(t-1)}\right)^{\mathsf{T}}, & 2 &\le t \le 3,\ 1 \le i \le n,
\end{aligned}$$

denoting the vectors containing the explanatory, mediator, and response variables prior to time t, respectively. Let

$$H_{it} = \left\{\widetilde{\mathbf{x}}_{it}, \widetilde{\mathbf{z}}_{it}, \widetilde{\mathbf{y}}_{it}; 2 \le t \le 3\right\}.$$

In the above, H_{it} contains all the observed information prior to time t. It follows from the posited assumptions of MAR that

$$\pi_{it} = Pr\left(r_{it}=1 \mid \mathbf{x}_i, \mathbf{z}_i, \mathbf{y}_i\right) = Pr\left(r_{it}=1 \mid H_{it}\right), \quad 2 \le t \le 3.$$

Thus, under MAR, the independence of r_{it} with $\mathbf{x}_i$, $\mathbf{z}_i$ and $\mathbf{y}_i$ under MCAR is replaced by a weaker version of conditional independence; r_{it} is independent of (x_{it}, z_{it}, y_{it}) given H_{it}.

To estimate π_{it}, let $p_{it} = \Pr(r_{it} = 1 \mid r_i(t-1) = 1, H_{it})$, denoting the one-step transition probability for observing data from time $t-1$ to t. We can model p_{it} using logistic regression:

$$\begin{aligned}
\operatorname{logit}\left(p_{it}\left(\eta_t\right)\right) &= \operatorname{logit}\left[Pr\left(r_{it}=1 \mid r_{i(t-1)}=1, H_{it}\right)\right] \\
&= \eta_{0t} + \boldsymbol{\eta}_{xt}^{\mathsf{T}}\widetilde{\mathbf{x}}_{it} + \boldsymbol{\eta}_{zt}^{\mathsf{T}}\widetilde{\mathbf{z}}_{it} + \boldsymbol{\eta}_{yt}^{\mathsf{T}}\widetilde{\mathbf{y}}_{it}, \quad 2 \le t \le 3,
\end{aligned} \quad (15)$$

where $\eta_t = \left(\eta_{0t}, \eta_{xt}^{\mathsf{T}}, \eta_{zt}^{\mathsf{T}}, \eta_{yt}^{\mathsf{T}}\right)^{\mathsf{T}}$ are the parameters for the logistic model above. It is readily shown that under MMDP

$$\pi_{it}(\eta) = p_{it}\, Pr\left(r_{i(t-1)} = 1 \mid H_{i(t-1)}\right) = \prod_{s=2}^{t} p_{is}(\eta_s), \quad 2 \le t \le 3,\ 1 \le i \le n, \quad (16)$$

where $\eta = (\eta_2^{\mathsf{T}}, \eta_3^{\mathsf{T}})^{\mathsf{T}}$ contains the collection of all parameters for modeling $\pi_{it}(\boldsymbol{\eta})$ ($t = 2, 3$). Thus, we can estimate π_{it} from the estimates of p_{it} using the relationship in (16).

To estimate η, we can use either maximum likelihood or estimating equations. For example, when using maximum likelihood, we estimate $\hat{\eta}$ as the solution to the score-based estimating equations:

$$\mathbf{Q}_n(\gamma) = \sum_{i=1}^{n} (\mathbf{Q}_{i2}^{\mathsf{T}}, \mathbf{Q}_{i3}^{\mathsf{T}})^{\mathsf{T}} = \mathbf{0}, \quad (17)$$

where

$$\mathbf{Q}_{it} = \frac{\partial}{\partial \eta_t}\left\{ r_{i(t-1)}\left[r_{it}\, log\,(p_{it}) + (1 - r_{it})\, log\,(1 - p_{it}) \right] \right\}, \quad (18)$$
$$2 \le t \le 3,\ i = 1, 2, \ldots, n,$$

With an estimated π_{it}, we can develop a set of weighted GEE II (WGEE II) for inference about θ. Consider first the FRM in (7) and define a weighting matrix as follows:

$$\boldsymbol{\Delta}_i = \begin{pmatrix} \frac{r_{i3}}{\pi_{i3}} & 0 & 0 & 0 & 0 \\ 0 & \frac{r_{i2}}{\pi_{i2}} & 0 & 0 & 0 \\ 0 & 0 & \frac{r_{i3}}{\pi_{i3}} & 0 & 0 \\ 0 & 0 & 0 & \frac{r_{i3}}{\pi_{i3}} & 0 \\ 0 & 0 & 0 & 0 & \frac{r_{i2}}{\pi_{i2}} \end{pmatrix}, \quad 1 \le i \le n. \quad (19)$$

Now consider the following WGEE II:

$$\mathbf{w}_n(\theta) = \sum_{i=1}^{n} \mathbf{w}_{ni} = \sum_{i=1}^{n} D_i V_i^{-1} \boldsymbol{\Delta}_i S_i = \mathbf{0}, \quad (20)$$

where Δ_i is given in (19), and D_i, V_i, and S_i are defined the same as in the equations in (12) except for the redefined longitudinal response $\mathbf{f}_i$ and mean function $\mathbf{h}_i$. As in the GEE II case, V_i may be computed under the normal distribution assumption. Finally, to compute an estimate of θ from (20), we must substitute an estimate of γ such as the one defined by (17) and (18).

The estimate $\hat{\theta}$ obtained from solving the WGEE II in (20) has nice asymptotic properties, akin to GEE II, as summarized in a theorem below, with a sketch of the proof given in Appendix A.

Theorem 1. *Under mild regularity conditions,*

1. $\hat{\theta}$ *is consistent.*

2. $\hat{\theta}$ is asymptotically normal with asymptotic variance given by

$$
\begin{aligned}
\Sigma_\theta &= B^{-1}\left(\Sigma_U + \Phi\right) B^{-\intercal}, \qquad B = E\left(D_i V_i^{-1} \boldsymbol{\Delta}_i D_i^\intercal\right), \\
\Sigma_U &= E\left(D_i V_i^{-1} \boldsymbol{\Delta}_i S_i S_i^\intercal \boldsymbol{\Delta}_i V_i^{-1} D_i^\intercal\right), \qquad \Phi = -CH^{-T}C^\intercal - F - F^\intercal, \\
C &= E\left[\frac{\partial}{\partial \eta}\left(D_i V_i^{-1} \boldsymbol{\Delta}_i S_i\right)\right]^\intercal, \qquad H = E\left(\frac{\partial}{\partial \eta} \mathbf{Q}_{ni}\right)^\intercal, \\
F &= E\left(D_i V_i^{-1} \boldsymbol{\Delta}_i S_i \mathbf{Q}_{ni}^\intercal H^{-\intercal} C^\intercal\right).
\end{aligned}
\tag{21}
$$

A consistent estimate of Σ_θ is given by

$$
\begin{aligned}
\hat{\Sigma}_\theta &= \hat{B}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{D}_i \hat{V}_i^{-1} \hat{\boldsymbol{\Delta}}_i \hat{S}_i \hat{S}_i^\intercal \hat{\boldsymbol{\Delta}}_i \hat{V}_i^{-1} \hat{D}_i^\intercal\right)\hat{B}^{-\intercal}, \quad \hat{B} = \frac{1}{n}\sum_{i=1}^{n}\hat{D}_i \hat{V}_i^{-1} \hat{\boldsymbol{\Delta}}_i \hat{D}_i^\intercal, \\
\hat{C} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \eta}\left(\hat{D}_i \hat{V}_i^{-1} \hat{\boldsymbol{\Delta}}_i \hat{S}_i\right), \quad \hat{F}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{D}_i \hat{V}_i^{-1} \hat{\boldsymbol{\Delta}}_i \hat{S}_i \hat{\mathbf{Q}}_{ni}^\intercal\right)\hat{H}^{-\intercal}\hat{C}^\intercal,
\end{aligned}
\tag{22}
$$

where $\hat{A}_i$ denotes the corresponding quantity with θ replaced by $\hat{\theta}$. Note that the asymptotic variance in (21) contains a modifying term $B^{-1}\Phi B^{-\intercal}$ to account for the sampling variability in the estimated $\hat{\eta}$. By substituting the first 2×2 submatrix of $_i$ defined in (19) in place of $_i$ in (20), we can apply the resulting WGEE II to provide inference about the θ for the FRM in (10).

4.3. A Score Test

As Wald-type tests are typically anticonservative (Daniels & Kass, 2001; Pan, 2001; Yu et al., 2013), score statistics are often used as an alternative to reduce bias, especially for type I error rates for small to moderate samples. We develop a score statistic based on the WGEE II, which is asymptotically equivalent to theWald, but is often more accurate for small and moderate samples.

Let $\theta = \left(\theta_{(1)}^\intercal, \theta_{(2)}^\intercal\right)^\intercal$, with p and q denoting the dimension of $\theta_{(1)}$ and $\theta_{(2)}$. Consider testing the null $H_0 : \theta_{(2)} = \theta_{(20)}$, with $\theta_{(20)}$ denoting a vector of known constants. Let

$$
\begin{aligned}
D_i &= \begin{pmatrix} \frac{\partial \mathbf{h}(\theta)}{\partial \theta_{(1)}} \\ \frac{\partial \mathbf{h}(\theta)}{\partial \theta_{(2)}} \end{pmatrix} = \begin{pmatrix} D_{i(1)} \\ D_{i(2)} \end{pmatrix}, \\
D_i V_i^{-1} \boldsymbol{\Delta}_i S_i &= \begin{pmatrix} D_{i(1)} V_i^{-1} \boldsymbol{\Delta}_i S_i \\ D_{i(2)} V_i^{-1} \boldsymbol{\Delta}_i S_i \end{pmatrix}, \\
\mathbf{w}_{n(1)}(\theta) &= \frac{1}{n}\sum_{i=1}^{n}D_{i(1)} V_i^{-1} \boldsymbol{\Delta}_i S_i, \qquad \mathbf{w}_{n(2)}(\theta) = \frac{1}{n}\sum_{i=1}^{n}D_{i(2)} V_i^{-1} \boldsymbol{\Delta}_i S_i.
\end{aligned}
\tag{23}
$$

Denote by $\widetilde{\theta}_{(1)}$ the estimate of $\theta_{(1)}$ from solving the following reduced WGEE II:

$$
\mathbf{w}_{n(1)}\left(\theta_{(1)}, \theta_{(20)}\right) = \frac{1}{n}\sum_{i=1}^{n}D_{i(1)} V_i^{-1} \boldsymbol{\Delta}_i S_i = \mathbf{0}. \tag{24}
$$

Note that only $\theta_{(1)}$ is unknown in the equations above.

To define the score statistic, let

$$
\begin{aligned}
\widetilde{\theta} &= \begin{pmatrix} \widetilde{\theta}_{(1)} \\ \theta_{(20)} \end{pmatrix}, & B &= E\left(D_i V_i^{-1} \Delta_i D_i\right) = \begin{pmatrix} B_{11} & B_{12} \\ B_{12}^\intercal & B_{22} \end{pmatrix}, \\
G &= \begin{pmatrix} -B_{21} B_{11}^{-1} & \mathbf{I}_p \end{pmatrix}, & \Sigma_{(2)} &= G \Sigma_\theta^{(-B)} G^\intercal,
\end{aligned}
\tag{25}
$$

where $\mathbf{I}_p$ designates the $p \times p$ identity matrix, $\boldsymbol{B}_{11}$ denotes the $p \times p$ submatrix, B_{12} the $p \times q$ submatrix, and B_{22} the $q \times q$ submatrix from the partitioned $(p + q) \times (p + q)$ matrix B defined in (21), and $\Sigma_\theta^{(-B)} = B \Sigma_\theta B^\intercal$, with Σ_θ defined in (21). Let

$$
\widetilde{\mathbf{w}}_{n(2)} = \widetilde{\mathbf{w}}_{n(2)}^\intercal \left(\left(\widetilde{\theta}_{(1)}, \theta_{(20)}\right)\right), \qquad \Sigma_{(2)}^{-1} = \widetilde{\Sigma}_{(2)}^{-1}\left(\left(\widetilde{\theta}_{(1)}, \theta_{(20)}\right)\right),
$$

i.e., the quantities of $\mathbf{w}_{n(2)}^\intercal$ and $\Sigma_{(2)}$ with $\boldsymbol{\theta}$ substituted by $\widetilde{\theta}$.

Consider the score test statistic, $T_s\left(\left(\widetilde{\theta}_{(1)}, \theta_{(20)}\right)\right) = n \widetilde{\mathbf{w}}_{n(2)}^\intercal \widetilde{\Sigma}_{(2)}^{-1} \widetilde{\mathbf{w}}_{n(2)}^\intercal$. As asserted by Theorem 2, this statistic has an asymptotic (central) χ_q^2 distribution with q degrees of freedom, i.e.,

$$
T_s\left(\left(\widetilde{\theta}_{(1)}, \theta_{(20)}\right)\right) = n \widetilde{\mathbf{w}}_{n(2)}^\intercal \widetilde{\Sigma}_{(2)}^{-1} \widetilde{\mathbf{w}}_{n(2)}^\intercal \to d \chi_q^2. \tag{26}
$$

Theorem 2. *Under mild regularity conditions, if $H_0 : \theta_{(2)} = \theta_{(20)}$ holds true, the score test $T_s\left(\left(\widetilde{\theta}_{(1)}, \widetilde{\theta}_{(20)}\right)\right)$ has the asymptotic distribution in* (26). *A sketch of the proof is given in* Appendix B.

5. Applications

In this section, we illustrate the methodology with both simulated and real study data. Since only $\boldsymbol{\theta} = (\gamma_0, \gamma_{zy}, \gamma_{xy}, \beta_0, \beta_{xz})^\intercal$ is of primary interest for mediation analysis, we focus on the simplified FRM-based SEM in (10) for all simulated and real study data, except for the complete data case of the simulation study in which we also apply the FRM in (7) to estimate the variance parameters in addition to $\boldsymbol{\theta}$. We start with the simulation study.

5.1. Simulation Study

We carried out a series of simulation studies to examine the performance of the proposed approach with different sample sizes. Since the cross-sectional data is a special case of the longitudinal setting under complete data, we performed the simulations for the latter only in the order of no missing data, and missing data under MCAR and the MAR. We also include a power analysis.

We report results for sample size $n = 50$, 100, and 500 with complete data. Since we anticipate potential bias in the ML approach with missing data under MAR in violation of parametric assumptions, according to theoretical results, we define our large sample size in the missing data simulation as $n = 2000$ in place of $n = 500$ to evaluate how this bias

continues to increase with sample size. All simulations were performed based on a Monte Carlo sample size of 1,000, with data generated and FRM analysis performed using the R software (R Development Core Team, 2008).We assumed a continuous x_{i1} for the simulation study in the paper. However, the proposed approach works for continuous and discrete x_{i1}.

We briefly outline the results of the simulation studies with complete data using the SEM for the mediation model in (1) in this section, before a more in depth treatment of the results with missing data. The FRM in (7) showed almost no bias with complete data and normal errors. The simplified FRM in (10) performed similarly well to the existing methods (ML, MLR, WLS, and GLS) with complete data and normal errors.

We also performed the same simulations for the simplified FRM in (10) under nonnormal error terms, using the t-distributed with 3 degrees of freedom and chi-squared with 1 degree of freedom (scaled to have mean 0 and variance 1) error terms, i.e., $\epsilon_{zi}, \epsilon_{yi} \sim F(0, 1)$, where $F(0, 1)$ is either a t distribution or a rescaled chi-square with 1 degree of freedom. The FRM-based approach once again showed almost no bias and performed similarly to another robust approach, ML estimation with robust asymptotic standard errors (Satorra & Bentler, 1994, 1998) performed using the MLR option in MPlus (Muthén & Muthén, 1998–2010), under all sample sizes. Although ML estimation in theory does not provide consistent estimates for nonnormal data, it performed well for these two particular nonnormal distributions, since we failed to observe large bias in the ML point estimates. We did observe slightly smaller estimates for the asymptotic standard errors under all sample sizes using our FRM-based SEM and slightly different Type I errors compared to ML, although these differences were almost trivial at $n = 500$. Another distribution-free approach, WLS, showed some potential convergence issues with nonnormal data, given that some of the Monte Carlo runs under the t-distributed errors with 3 degrees of freedom lead to inflated estimates for the asymptotic standard errors for the estimates of $\boldsymbol{\gamma} = (\gamma_0, \gamma_{zy}, \gamma_{xy})^\top$ at $n = 100$ and at $n = 500$ when compared to FRM, MLR, ML, and GLS.

5.1.1. Missing Data—We considered the SEM for the longitudinal mediation model with central t-distributed random error terms with 3 degrees of freedom:

$$
\begin{aligned}
x_{i1} &= \mu_x + g_i + \epsilon_{xi1}, & z_{i1} &= \beta_{01} - g_i + f_i + \epsilon_{zi1}, \\
z_{i2} &= \beta_{02} + \beta_{12} x_{i1} + c_i + f_i + \epsilon_{zi2}, & y_{i2} &= \gamma_{02} - c_i + d_i + \epsilon_{yi2}, \\
y_{i3} &= \gamma_{03} + \gamma_{13} x_{i1} + \gamma_{23} z_{i2} + d_i + \epsilon_{yi3}, \\
\epsilon_{xi1} &= \epsilon_{zi1} = \epsilon_{zi2} = \epsilon_{yi2} = \epsilon_{yi3} \sim t_3 \\
g_i &\sim N\left(0, \sigma_{\mathbf{g}}^2\right), f_i \sim N\left(0, \sigma_{\mathbf{f}}^2\right), c_i \sim N\left(0, \sigma_{\mathbf{c}}^2\right), d_i \sim N\left(0, \sigma_{\mathbf{d}}^2\right).
\end{aligned}
\tag{27}
$$

Assuming no missing data at baseline $t = 1$, we simulated missing responses under MCAR and MAR with about 15 % (30 %) missing data at $t = 2$ (3). We also included random effects g_i, f_i, c_i, and d_i in our model for the between subject effects, linking observations of measures from previous time points to current time points, a common occurrence in real world data. Since the error terms are t-distributed, the joint normal distribution assumption is not met in the presence of missing data following MAR. Also, unlike the case in modeling a single response over time as in traditional longitudinal data analysis, the MAR mechanism

within the current context of mediation analysis is more complex, since missingness can occur to either or both of endogenous variables z_{it} and y_{it}.

Note that the random effects in (27) were used to create correlated repeated measures of both within- and between-variables. We could have used other approaches such as the copula method (Nelson, 2006). However, the random-effect approach allowed us to have better control about the relationships both within the same and across different variables.

We discussed how to model the missing data mechanism under MMDP in Section 4.2. In the simulation model in (27), the repeated measures of the three variables were linked together using random effects and thus missingness of any of the variables could be made to depend on itself or any of the other variables at the previous time points. For convenience and without the loss of generality, we considered a relatively simple Markov structure for the missing value indicator r_{it}, with the missingness of z_{i2} (y_{i3}) depending only on its predecessor z_{i1} (y_{i2}), with the one-step transition probability p_{it} in (15) taking the following form:

$$
\begin{aligned}
\text{logit}\,(p_{i2}) &= \text{logit}\,(Pr\,(r_{i2}=1 \mid r_{i1}=1, H_{i2})) = \eta_{02} + \eta_{z2} z_{i1}, \\
\text{logit}\,(p_{i3}) &= \text{logit}\,(Pr\,(r_{i3}=1 \mid r_{i2}=1, H_{i3})) = \eta_{03} + \eta_{y3} y_{i2}.
\end{aligned}
\tag{28}
$$

We set $\eta_{z2} = \eta_{y3} = 0.4$ ($\eta_{z2} = \eta_{y3} = 0$) for MAR (MCAR), and then selected values of η_{0t} to yield about 15 % and 30 % missing data at $t = 2$ and 3 under MAR (MCAR). We considered the simplified FRM-based SEM in (10), but simulated (x_{it}, z_{it}, y_{it}) over all three assessment times to enable us to model the missing data.

We fixed σ_g^2, σ_f^2, σ_c^2 and σ_d^2 so that, approximately:

$$
\begin{aligned}
\text{Corr}\,(x_{i1}, z_{i1}) &= -0.30, & \text{Corr}\,(z_{i1}, z_{i2}) &= 0.25, \\
\text{Corr}\,(z_{i2}, y_{i2}) &= -0.20, & \text{Corr}\,(y_{i2}, y_{i3}) &= 0.25.
\end{aligned}
$$

Therefore, in this simulation we chose a scenario using random effects to link repeated measures and a special case of our missing data model to illustrate the type of bias we might observe in practice with a moderate amount of missing data.

For the simulation study, we set $\theta = (\gamma_0, \gamma_{zy}, \gamma_{xy}, \beta_0, \beta_{xz})^\mathsf{T} = (1, 1, 0, 1, 1)^\mathsf{T}$. Since $\gamma_{xy} = 0$, we were able to examine the performance of the FRM when used to test the null hypothesis of full mediation. Given the correlations and distribution assumptions in (27), the outcomes, (x_{i1}, z_{i2}, y_{i3}), were readily simulated from (27). The data generated was fit by the FRM-based SEM in (7).

To examine type I error rates, we considered the null of full mediation, H_0: $\gamma_{xy} = 0$, for the parameter of primary interest, and computed the Wald statistic, $Q_n = n\hat{\sigma}_{\gamma_{xy}}^{-2} \hat{\gamma}_{xy}^2$, where $\hat{\sigma}_{\gamma_{xy}}^2$ denotes the element of the estimated asymptotic variance $\hat{\Sigma}_\theta$ corresponding to $\hat{\gamma}_{xy}$. Let $Q_n^{(k)}$ denote the value of this statistic from the kth MC simulation ($1 \leq k \leq 1000$). The type I error

rate for testing H_0 was estimated by: $\hat{\alpha} = \frac{1}{1000} \sum_{k=1}^{1000} I_{\{Q_n^{(k)} \geq q_{1,0.95}\}}$, with $q_{1,0.95}$ denoting the 95th percentile of the central χ_1^2 with one degree of freedom.

Since the Wald statistic is often anticonservative (Yu et al., 2013), we also applied the score test statistic in Section 4.3 to see how this alternative would improve the Wald within the current setting. For testing the null $H_0 : \gamma_{xy} = 0$, we partitioned $\boldsymbol{\theta}$ as $\theta = \left(\theta_{(1)}^{\mathsf{T}}, \theta_{(2)} \right)^{\mathsf{T}}$. with $\boldsymbol{\theta}_{(1)}$ $= (\gamma_{zy}, \beta_{xz})^{\mathsf{T}}$ and $\theta_{(2)} = \gamma_{xy}$. Under $H_0 : \theta_{(2)} = 0$, the score statistic $T_s\left(\tilde{\theta}_{(1)}, 0 \right)$ in (26) has an asymptotic χ_1^2 distribution. The type I error rate for the score test was calculated similarly as in the Wald by $\hat{\alpha} = \frac{1}{1000} \sum_{k=1}^{1000} I_{\{T_s^{(k)} \geq q_{1,0.95}\}}$, where $T_s^{(k)}$ denotes the value of the statistic $T_s\left(\tilde{\theta}_{(1)}, 0 \right)$ from the kth MC simulation ($1 \leq k \leq 1000$).

The estimate of $\boldsymbol{\theta}$ was obtained from the WGEE II. Both GLS and WLS perform listwise deletion and as a result do not provide valid inference in the presence of missing data. Therefore, it was unnecessary to include these estimates in the simulation studies. Since results under MCAR are quite similar to those under complete data for FRM and ML, we only report results for the MAR case. The full information maximum likelihood (FIML) approach provided by Mplus handled missing data under ML (Muthén & Muthén, 1998–2010).

Shown in Table 1 are the percent bias of parameter estimates, estimates of the average asymptotic standard errors over 1,000 MC replications and type I error rates for the SEM model in (27) under FRM and ML, with missingness modeled as in (28) for FRM. For comparison purposes, we also included the relative difference (RD) between the average asymptotic and "empirical" standard error estimates defined as

$$RD = \left| (\text{asymptotic} - \text{empirical}) \left/ \left(\frac{\text{asymptotic} + \text{empirical}}{2} \right) \right. \right|,$$

where $|\cdot|$ is the absolute value function.

The percent bias of parameter estimates for FRM was relatively small, and approached zero as the sample size increased. However, the estimates for ML showed a heavy bias, which stayed at about the same magnitude across the different samples. The Score type I error for FRM was close to or below 0.05 at all times, while the Wald type I error was inflated for $n = 50, 100$, but decreased to 0.05 at the large sample size of $n = 2000$. In contrast, the type I error for ML remained high and even increased to 0.159 at $n = 2000$, confirming that ML does not provide consistent estimates.

Figure 2 visually depicts the bias in ML compared to FRM. FRM, as the sample size increases, becomes even closer to a horizontal line representing the bias of parameter estimates (difference between the mean parameter estimates and population parameter estimates) for all the estimates of $\boldsymbol{\theta} = (\gamma_0, \gamma_{xy}, \gamma_{zy}, \beta_0, \beta_{xz})^{\mathsf{T}}$, while ML continues to show

bias with a curvilinear pattern, under and overestimating the different parameters. We see that the size of the confidence intervals for the estimates whether by FRM or ML are all decreasing with sample size, with the difference that ML estimates are off target. Additionally, we performed the same missing data simulation using MLR in place of ML. Since the percent bias of parameters estimates will be the same between the two methods, both are problematic. Similar to the case of complete data, for $n = 50, 100$, the asymptotic standard errors were slightly smaller for MLR. However, the type I error was higher for MLR (Type I error = 0.092) at $n = 50$, and $n = 100$. (Type I error = 0.69). Again, however, at the large sample, in this case $n = 2000$, there was no distinct difference between the two methods of MLR (Type I error = 0.162) and ML.

5.1.2. Power Analysis—Shown in Table 2 are the power estimates under sample sizes $n = 50, 100$, and 500 for the Wald test for the different methods generated under full mediation, $H_0 : \gamma_{xy} = 0$ vs. $H_a : \gamma_{xy} = 0.5$, considered for the SEM model in (1), with complete data and normal errors as well as missing data following MAR modeled by (27) under correlated repeated measures and t-distributed errors. Under complete data, the Score test showed that the proposed approach had slightly lower power, but still performed well. Note that both the Wald test for FRM and WLS showed higher power than ML for $n = 50, 100$, but these were likely the result of bias in the estimates due to small to moderate sample sizes, since ML was (asymptotically) the most powerful in the simulation setting with normal errors and complete data.

Since we already established a bias in ML with missing data under MAR with nonnormal error terms, we wanted to see how FRM would perform under missing data in terms of detecting mediation effects. Since only FRM yielded valid inference in this case, the results in the table for the large sample size $n = 500$ indicated that ML had an upward bias. We have noted during the simulation study with missing data for the SEM model in (27) that the bias in ML is increasing with sample size, so this increase in power may lead a researcher to believe even more strongly in biased results.

As in the complete data case, the Wald test for FRM again showed some upward bias, especially for the small and moderate sample sizes $n = 50, 100$. Note that we did not include power estimates for GLS and WLS since these methods perform listwise deletion, which not only yielded inefficient estimates under missing data, but also invalid inference under MAR.

5.2. Real Study Data

The real study data is shown to complement the simulation results by showing how the methods may lead us to different estimates and conclusions when presented with nonnormal, skewed complete data and missing data under MAR.

5.2.1. Child Resilience Example with Complete Data—To illustrate the approach to real study data, we applied FRM to a longitudinal study known as the Child Resilience Project (Wyman, Cross, Brown, Yu, & Tu, 2010). This project is ongoing, with 401 students from first up to third grade in five Rochester City School District elementary schools. Enrollment began in Fall 2006, with data collection for the final cohort to be completed by June 2011. The study examines how children with a higher risk of developing behavioral

problems with a mentor socially improve compared to the control and lower risk children over periods of 6 and 18 months.

In this mediation analysis, both the mediator and response were focused on helping children to manage challenging emotions—emotion self-regulation, self-reported verbal, declarative knowledge of the skills the child is learning in the Resilience Project at 6 months. We examined what role a potential mediator plays in a cause and effect relationship between the treatment and the child's self-initiated demonstration of skills he/she is learning at 6 months. Thus, we have longitudinal data with two assessment times, baseline and 6 months, with the mediator hypothesized to occur before the outcome. There were 253 subjects with these three measures forming a complete dataset. Shown in Table 3 is a sample of our longitudinal data set for the three variables of interest.

The treatment is a binary indicator as children either had a mentor or no mentor. In the hypothesis of interest, the treatment would be expected to predict a higher demonstration of skills, which would indicate that the children receiving a mentor improved their social skills over time. The distributions of both the verbal, declarative knowledge of skills, and demonstration of skills were skewed as shown in Figure 3.

By identifying the treatment condition as x_{i1}, the verbal, declarative knowledge of the skills at 6 months as z_{i2}, and the demonstration of skills at 6 months as y_{i3}, we modeled the mediation process of interest using the SEM in (1).

Shown in Figure 4 are the path diagrams for the direct linear model and full mediation model. Since the path for the linear regression was significant, we looked to the mediation model to see how the mediator would effect the relationship between the treatment and demonstration of skills.

Shown in Table 4 are the estimates of θ, standard errors, and type I errors for this mediation model obtained from FRM and ML. Since the results for ML and GLS were practically the same, we only included the ML estimates in the table. WLS was unable to yield estimates for this highly skewed data, showing the same potential convergence problems as for some of the Monte Carlo runs with the t-distributed with 3 degrees of freedom error terms in our simulation studies. From the table, we see that the estimates for FRM and ML were practically the same. This is not surprising, since, as noted in our simulation study, ML would still produce consistent estimates in the presence of nonnormal error under complete data. However, the standard error was significantly lower for the FRM estimate of γ_0, and the standard errors were also lower for the FRM estimates of $\boldsymbol{\beta} = (\beta_0, \beta_{xz})^\top$. Although the FRM and ML estimates of γ_{xy}, the parameter of primary interest to the mediation hypothesis, were identical to the three decimal places shown in the table, the slightly higher standard error and type I error of FRM may be the most accurate. Since we hypothesized (both conceptually as well as based on other related findings) that a demonstration of skills was positively associated with treatment in this model, the estimates of γ_{xy} confirmed this relationship.

The relationship was significant in the linear model, and the magnitude of the relationship has decreased, while the p-value has increased in the mediation model. Therefore, verbal,

declarative knowledge of skills at 6 months may be a partial mediator. It is important to note that all parameters $\theta = (\gamma_0, \gamma_{xy}, \gamma_{zy}, \beta_0, \beta_{xz})^\top$ were highly significant for FRM by both Wald and Score Type I errors (p 0.005). However, for ML the γ_0 parameter was not significant ($p = 0.053$).

5.2.2. Child Resilience Example with Missing Data—As mentioned in Section 5.2.1, there were 401 students in the child resilience project. We had full information on whether each child received the treatment at baseline. However, there were more missing observations for demonstration of skills at 18 months, as now only 164 students were observed. A longitudinal mediation analysis under missing data, using the same three variables as in Section 5.2.1, over 18 months will help us understand if the treatment is effective over a longer time period, accounting for all 401 children in the study.

When modeling for the missing data, we had 253 observations for the mediator, declarative knowledge of the skills at 6 months, as in the complete data example. Shown in Table 5 is a sample of our longitudinal data set for the three variables of interest.

We expanded our mediation analysis to an 18-month time period with missing data by identifying the treatment condition as x_{i1}, the verbal, declarative knowledge of the skills at 6 months as z_{i2}, and the demonstration of skills at 18 months as y_{i3}, and model the mediation process of interest using the SEM in (1), as shown in Figure 5.

We had a high percentage of (37 %/59 %) missing data at $t = 2(3)$ and modeled our missing data using the following logistic model:

$$\text{logit}\,(p_{i2}) = \eta_{02} + \eta_{x1} x_{i1}, \qquad \text{logit}\,(p_{i3}) = \eta_{03} + \eta_{z2} z_{i2}, \quad (29)$$

$$p_{it} = Pr\left(r_{it} = 1 \mid r_{i(t-1)} = 1, H_{it}\right). \quad (30)$$

This is a special case of our missing data model from Section 4.2 in which we are only observing x_{i1}, z_{i2}, and y_{i3} and building our missing data models with only observed data at the previous time point from x_{i1} and z_{i2}. We estimated the parameters in R with the glm function. Since we modeled our missing data at $t = 2$ based on the treatment information at baseline, we used all 401 observations.

Shown in Table 6 are the estimates of $\eta = (\eta_{02}, \eta_{x1}, \eta_{03}, \eta_{z2})$ for the missing data model in (29). The p-value for η_{z2} was significant, indicating a MAR mechanism for the missing data at time 3. Since the p-value for η_{x1} was not significant, missing data at time 2 was MCAR and we would expect no bias for the estimates of $\beta = (\beta_0, \beta_{xz})^\top$ in ML according to our simulation results in Section 5.1.1. However, based on our simulation results in Table 1, we expect to see a bias for the estimates of the $\gamma = (\gamma_0, \gamma_{xy}, \gamma_{zy})^\top$ parameters in ML. Subsequently, if we assumed a MAR mechanism for the missing data at time 2 and a MCAR mechanism for the missing data at time 3, we would expect bias for the estimates of $\beta = (\beta_0, \beta_{xz})^\top$ at time 2 but not for $\gamma = (\gamma_0, \gamma_{xy}, \gamma_{zy})^\top$ at time 3 in ML. In the hypothesis of interest, the treatment would be expected to predict a higher demonstration of skills at 18 months,

which would indicate that the children receiving a mentor improved their social skills over time.

Shown in Table 7 are the estimates of θ and associated standard errors and type I errors for this mediation model obtained from FRM and ML. From the table, we see that the estimates for FRM and ML were practically the same for the $\boldsymbol{\beta} = (\beta_0, \beta_{xz})^{\top}$ parameters, but different for $\boldsymbol{\gamma} = (\gamma_0, \gamma_{xy}, \gamma_{zy})^{\top}$. In the three significant estimates with the same or similar results for FRM and ML, $(\gamma_0, \beta_0, \beta_{xz})^{\top}$, FRM had a smaller standard error. We saw from the simulation for longitudinal missing data in Table 3 that ML would produce a value of γ_{xy} biased less in magnitude than the true estimate. This appeared true again as the FRM estimate was higher in magnitude, confirming that the treatment predicted a higher demonstration of skills at 18 months. The parameter γ_{zy} was not significant for either FRM or ML in this model ($p >$ 0.421 for the Wald test in both FRM and ML), implying a nonsignificant indirect effect in this mediation analysis.

6. Discussion

Mediation analysis is a critical component of many studies in biomedical, psychosocial, and related services research to investigate the causal mechanism of interventions. The traditional regression paradigm is ill-suited for modeling such multidimensional causal relationships, because of the complex relationships among different variables and many different roles a variable may play. Although the structural equation model (SEM) provides an ideal conceptual framework for the dynamic role a mediator plays in channeling the effect of a causal agent to the outcome of interest, the various restrictions imposed by the available inference methods have hindered its wide use, and limited the utility of the information provided in real studies. Namely, in real studies, commonly we come across missing data, where making the MAR assumption is reasonable and parametric assumptions may be violated.

Great progress has been made in the analysis of causal relationships, particularly in the area of distribution-free models for longitudinal data analysis, over the past few decades. However, little of this progress applies directly to SEM and its application to modeling mediation relationships within a longitudinal setting. One major block is the nonlinear nature of SEM in its parameters, which in general requires second- and even higher-order moments to identify model parameters. As only the first-order moment is modeled in the standard regression paradigm, addressing the limitations of SEM requires a paradigm shift by breaking the tradition in current regression analysis and creating a new framework for modeling nonlinear relationships and higher-order moments.

By taking advantage of the functional response models (FRM), we have developed a new, robust approach to systematically address the limitations of SEM as it applies to mediation analysis. This class of FRM-based SEM requires no parametric models for the data distribution and provides valid inference for longitudinal mediation hypotheses under the two most popular missing data mechanisms, missing completely at random (MCAR) and missing at random (MAR).

The results from the simulation studies indicate that the new approach performed well under both complete and missing data for small and moderate sample sizes. Under complete data, the FRM-based SEM is nearly on par with the parametric ML in efficiency. Thus, the loss of efficiency seems to be negligible for practical purposes. On the other hand, the FRM-based mediation model shines when the missing data follow MAR with nonnormal errors, since it is the only model to provide robust estimates and valid inference.

The validity of WGEE II relies on a correct model for the missingness. If this model is misspecified, WGEE II estimates may be biased. Work is currently underway to investigate the possibility to extend the concept of double robust estimate in traditional distribution-free regression models to the current context of FRM-based mediation model (Browne, 1974; Lu et al., 2009; Satorra, 2002). Currently, in progress are also extensions of the approach to noncontinuous outcomes such as binary.

Acknowledgements

Appendix A. Proof of Theorem 1

Let $G_i = D_i V_i^{-1}$ and $\pi_i = (\pi_{i1}, \dots, \pi_{im1})^\top$. Then $\mathbf{w}_n = \frac{1}{n}\sum_{i=1}^{n} G_i \boldsymbol{\Delta}_i S_i$, with $G_i \;_i S_i = G_i(\mathbf{x}_i, \boldsymbol{\theta})\;_i(\mathbf{r}_i, \pi_i, \gamma) S_i(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\theta})$. It follows from the iterated conditional expectation that $E(G_i \;_i S_i)$ $= E[G_i S_i E(\;_i \mid \mathbf{r}_i, \mathbf{y}_i, \mathbf{x}_i)]$. Since $E\left(\frac{r_{it}}{\pi_{it}} \;\middle|\; \mathbf{r}_i, \mathbf{y}_i, \mathbf{x}_i\right) = 1$, it follows that $E(G_i \;_i S_i) = E(G_i S_i) = 0$. Thus, the WGEE II is unbiased and the estimate $\hat{\theta}$ obtained as the solution to the equations is consistent.

Let $\hat{\gamma}$ be the solution to (17). By a Taylor expansion of the estimating equations in (17) and solving for $\hat{\gamma} - \gamma$, we obtain

$$\sqrt{n}\,(\hat{\gamma} - \gamma) = -H^{-1}\frac{\sqrt{n}}{n}\sum_{i=1}^{n}\mathbf{Q}_{ni} + \mathbf{o}_p(1), \quad \text{(A.1)}$$

where $\mathbf{o}_p(1)$ denotes the stochastic $\mathbf{o}(1)$ (Kowalski & Tu, 2007). Also, by applying a Taylor series expansion to the WGEE II in (20), we have

$$\begin{aligned}
\sqrt{n}\mathbf{w}_n &= -\left(\tfrac{\partial}{\partial\theta}\mathbf{w}_n\right)^\top \sqrt{n}\left(\hat{\theta} - \theta\right) - \left(\tfrac{\partial}{\partial\gamma}\mathbf{w}_n\right)^\top \sqrt{n}\,(\hat{\gamma} - \gamma) + \mathbf{o}_p(1) \\
&= \left(-\tfrac{\partial}{\partial\theta}\mathbf{w}_n\right)^{-\top} \sqrt{n}\left[\mathbf{w}_n + C\,(\hat{\gamma} - \gamma)\right] + \mathbf{o}_p(1).
\end{aligned}$$

It follows from (A.1) and (A.2) that

$$\sqrt{n}\left(\hat{\theta}-\theta\right)=\left(-\frac{\partial}{\partial\theta}\mathbf{w}_n\right)^{-\intercal}\frac{\sqrt{n}}{n}\sum_{i=1}^{n}\left(\mathbf{w}_{ni}-CH^{-1}\mathbf{Q}_{ni}\right)+\mathbf{o}_p\left(1\right). \quad \text{(A.2)}$$

Since

$$\frac{\partial}{\partial\theta}\mathbf{w}_n=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta}\Delta_i S_i\right)G_i^{\intercal}+\mathbf{o}_p\left(1\right)=\frac{1}{n}\sum_{i=1}^{n}D_i\Delta_i G_i^{\intercal}+\mathbf{o}_p\left(1\right)\to_p-B, \quad \text{(A.3)}$$

where $\to_p$ denotes convergence in probability, it follows from (A.2) and (A.3) that

$$\sqrt{n}\left(\hat{\theta}-\theta\right)=-B^{-\intercal}\frac{\sqrt{n}}{n}\sum_{i=1}^{n}\left(\mathbf{w}_{ni}-CH^{-1}\mathbf{Q}_{ni}\right)+\mathbf{o}_p\left(\frac{1}{2}\right). \quad \text{(A.4)}$$

By applying the central limit theorem and Slutsky's theorem (Kowalski & Tu, 2007) to (A.4), $\hat{\theta}$ is asymptotically normal with the asymptotic variance given by Σ_θ in (21).

Appendix B. Proof of Theorem 2

First, assume no missing data. Then $B=E\left(D_i V_i^{-1} D_i\right)$. By applying the law of large numbers,

$$\frac{\partial}{\partial\theta}\mathbf{w}_n\left(\theta\right)=\left(\begin{array}{cc}\frac{\partial}{\partial\theta_{(1)}}\mathbf{w}_{n(1)}\left(\theta\right) & \frac{\partial}{\partial\theta_{(1)}}\mathbf{w}_{n(2)}\left(\theta\right)\\\frac{\partial}{\partial\theta_{(2)}}\mathbf{w}_{n(2)}\left(\theta\right) & \frac{\partial}{\partial\theta_{(2)}}\mathbf{w}_{n(2)}\left(\theta\right)\end{array}\right)\to_p B=E\left(D_i V_i^{-1} D_i\right)=\left(\begin{array}{cc}B_{11} & B_{12}\\B_{12}^{\intercal} & B_{22}\end{array}\right). \quad \text{(B.1)}$$

It follows from a Taylor's series expansion and (B.1) that

$$\mathbf{0}=\mathbf{w}_{n(1)}\left(\tilde{\theta}_{(1)},\theta_{(20)}\right)=\mathbf{w}_{n(1)}\left(\theta\right)-\mathbf{B}_{11}^{-\intercal}\left(\tilde{\theta}_{(1)}-\theta_{(1)}\right)+\mathbf{o_p}\left(\mathbf{n}^{-\frac{1}{2}}\right).$$

Thus,

$$\tilde{\theta}_{(1)}-\theta_{(1)}=B_{11}^{-1}\mathbf{w}_{n(1)}\left(\theta\right)+\mathbf{o}_p\left(n^{-\frac{1}{2}}\right). \quad \text{(B.2)}$$

Similarly, since $B_{12}^{\intercal}=B_{21}$, we have

$$\begin{aligned}\mathbf{w}_{n(2)}\left(\tilde{\theta}_{(1)},\theta_{(20)}\right)&=\mathbf{w}_{n(2)}\left(\theta\right)-\left(\frac{\partial^{\intercal}}{\partial\theta_{(1)}}\mathbf{w}_{n(2)}\right)\left(\tilde{\theta}_{(1)}-\theta_{(1)}\right)+\mathbf{o}_p\left(n^{-\frac{1}{2}}\right)\\&=\mathbf{w}_{n(2)}\left(\theta\right)-B_{21}\left(\tilde{\theta}_{(1)}-\theta_{(1)}\right)+\mathbf{o}_p\left(n^{-\frac{1}{2}}\right).\end{aligned} \quad \text{(B.3)}$$

It follows from (B.2) and (B.3) that

$$\begin{aligned}\mathbf{w}_{n(2)}\left(\tilde{\theta}_{(1)},\theta_{(20)}\right)&=\mathbf{w}_{n(2)}\left(\theta\right)-B_{21}\left[B_{11}^{-1}\mathbf{w}_{n(1)}\left(\theta\right)+\mathbf{o}_p\left(n^{-\frac{1}{2}}\right)\right]+\mathbf{o}_p\left(n^{-\frac{1}{2}}\right)\\&=G\mathbf{w}_n\left(\theta\right)+\mathbf{o}_p\left(n^{-\frac{1}{2}}\right).\end{aligned}$$

By the central limit theorem,

$$\sqrt{n}\mathbf{w}_{n(2)}\left(\widetilde{\theta}_{(1)},\theta_{(20)}\right) = \sqrt{n}G\mathbf{w}_n\left(\theta\right) + \mathbf{o}_p\left(1\right) \rightarrow_d N\left(\mathbf{0},\boldsymbol{\Sigma}_{(\mathbf{2})} = \mathbf{G}\boldsymbol{\Sigma}_\theta\mathbf{G}^\intercal\right),\quad \text{(B.4)}$$

where G is defined in (25) and Σ_θ in (21). In the presence of missing data, $B = E\left(D_i V_i^{-1}\boldsymbol{\Delta}_i D_i\right)$ as defined in (21). By a similar argument, $\mathbf{w}_{n(2)}\left(\widetilde{\theta}_{(1)},\theta_{(20)}\right)$ has an asymptotic normal distribution, which implies that the score statistic $T_s\left(\left(\widetilde{\theta}_{(1)},\theta_{(20)}\right)\right)$ has the asymptotic χ_q^2 distribution.

References

Allison, PD. Sociological methodology. American Sociological Association; Washington: 1987. Estimation of linear models with incomplete data; p. 71-103.

Arbuckle, JL. IBM SPSS Amos 19 user's guide. Amos Development Corporation; Crawfordville: 1995–2010.

Baker LA, Fulker DW. Incomplete covariance matrices and LISREL. Data Analyst. 1983; 1:3–5.

Baron RM, Kenny DA. The moderator–mediator variable distinction in social psychological research: concept, strategic and statistical considerations. Journal of Personality and Social Psychology. 1986; 51:1173–1182. [PubMed: 3806354]

Bentler, PM. EQS 6 structural equations program manual. Multivariate Software, Inc.; Encino: 2006.

Bollen, KA. Structural equations with latent variables. Wiley; New York: 1989.

Bollen, KA.; Pearl, J. Eight myths about causality and structural equation models (Technical Report (R-393)). In: Morgan, S., editor. Handbook of causal analysis for social research. Springer; New York: 2012. UCLA Cognitive Systems Laboratory, Draft chapter for

Bollen KA, Stine R. Direct and indirect effects: classical and bootstrap estimates of variability. Sociological Methodology. 1990; 20:115–140.

Browne MW. Generalized least-squares estimators in the analysis of covariance structures. South African Statistical Journal. 1974; 8:1–24.

Clogg CC, Petkova E, Shihadeh ES. Statistical methods for analyzing collapsibility in regression models. Journal of Educational Statistics. 1992; 17(1):51–74.

Daniels MJ, Kass RE. Shrinkage estimators for covariance matrices. Biometrics. 2001; 57:1173–1184. [PubMed: 11764258]

Gunzler, D. Doctoral dissertation. 2011. A class of distribution-free models for longitudinal mediation analysis. Retrieved from ProQuest Dissertations and Theses Database (AAT 3478294)

Imai K, Keele L, Tingley D. A general approach to casual mediation analysis. Psychological Methods. 2010; 15(4):309–334. [PubMed: 20954780]

Joreskog, KG.; Sorbom, D. Lisrel 8 user's guide. 2nd ed.. Scientific Software; Lincolnwood: 1997.

Kowalski J, Powell J. Nonparametric inference for stochastic linear hypotheses: application to high-dimensional data. Biometrika. 2004; 91(2):393–408.

Kowalski, J.; Tu, XM. Modern applied U statistics. Wiley; New York: 2007.

Kowalski J, Pagano M, DeGruttola V. A nonparametric test of gene region heterogeneity associated with phenotype. Journal of the American Statistical Association. 2002; 97(458):398–408.

Kraemer H. How do risk factors work together?Mediators, moderators, and independent, overlapping, and proxy risk factors. The American Journal of Psychiatry. 2001; 158:848–856. [PubMed: 11384888]

Little, RJA.; Rubin, DB. Statistical analysis with missing data. Wiley; New York: 1987.

Lu N, Tang W, He H, Yu Q, Crits-Christoph P, Zhang H, Tu XM. On the impact of parametric assumptions and robust alternatives for longitudinal data analysis. Biometrical Journal. 2009; 51:627–643. [PubMed: 19688758]

Ma Y, Tang W, Feng C, Tu XM. Inference for kappas for longitudinal study data: applications to sexual health research. Biometrics. 2008; 64:781–789. [PubMed: 18047535]

Ma Y, Tang W, Yu Q, Tu XM. Modeling concordance correlation coefficient for longitudinal study data. Psychometrika. 2010; 75(1):99–119.

Maas CJ, Hox JJ. Robustness issues in multilevel regression. Statistica Neerlandica. 2004; 58:127–137.

MacKinnon, D. Introduction to statistical mediation analysis. Lawrence Erlbaum Associates; New York: 2008.

MacKinnon D, Fairchild A. Current directions in mediation analysis. Current Directions in Psychological Science. 2009; 18:16–20. [PubMed: 20157637]

McCullagh, P.; Nelder, JA. Generalized linear models. Chapman and Hall; London: 1989.

Muthén, LK.; Muthén, BO. Mplus user's guide. 6th ed.. Muthén & Muthén; Los Angeles: 1998–2010.

Nelsen, RB. An introduction to Copulas. Springer; New York: 2006.

Pan W. On the robust variance estimator in generalized estimating equations. Biometrika. 2001; 88:901–906.

Papadopoulos S, Amemiya Y. Correlated samples with fixed and nonnormal latent variables. The Annals of Statistics. 2005; 33:2732–2757.

Preacher, KJ.; Wichman, AL.; MacCallum, RC.; Briggs, NE. Latent growth curve modeling. Sage; Los Angeles: 2008.

Prentice RL, Zhao LP. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. Biometrics. 1991; 47:825–839. [PubMed: 1742441]

R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing; Vienna: 2008.

Reboussin BA, Liang KY. An estimating equations approach for the LISCOMP model. Psychometrika. 1998; 63:165–182.

Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association. 1995; 90:106–121.

Rothman, KJ.; Greenland, S. Modern epidemiology. Lippingcott Williams and Wilkins; Philadelphia: 1998.

Satorra A. Asymptotic robustness in multiple group linear-latent variable models. Econometric Theory. 2002; 18:297–312.

Satorra, A.; Bentler, PM. Corrections to test statistics and standard errors in covariance analysis. In: von Eye, A.; Clogg, CC., editors. Latent variable analysis: applications to developmental research. Sage; Newbury Park: 1994. p. 399-419.

Satorra, A.; Bentler, PM. Scaling corrections for chi-square statistics in covariance structure analysis; In 1988 proceedings of the business and economics statistics section of the American Statistical Association; 1998; p. 308-313.

Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models. Journal of the American Statistical Association. 1999; 94:1096–1120.

Sobel, ME. Asymptotic intervals for indirect effects in structural equations models. In: Leinhart, S., editor. Sociological methodology. Jossey-Bass; San Francisco: 1982. p. 290-312.

Tsiatis, AA. Semiparametric theory and missing data. Springer; New York: 2006.

Tu XM, Feng C, Kowalski J, Tang W, Wang H, Wan C, Ma Y. Correlation analysis for longitudinal data: applications to HIV and psychosocial research. Statistics in Medicine. 2007; 26:4116–4138. [PubMed: 17342700]

Van der Leeden, R.; Busing, F. First iteration versus IGLS/RIGLS estimates in two-level models: a Monte Carlo study with ML3. Department of Psychometrics and Research Methodology, Leiden University; Leiden: 1994.

Van der Leeden, R.; Busing, F.; Meijer, E. Applications of bootstrap methods for two-level models; Paper presented at the Multilevel Conference; Amsterdam, The Netherlands. 1997; Apr 1–2.

Werts CE, Rock DA, Grandy J. Confirmatory factor analysis applications: missing data problems and comparisons of path models between populations. Multivariate Behavioral Research. 1979; 14:199–213. [PubMed: 26815152]

Wyman PA, Cross W, Brown HC, Yu Q, Tu XM. Intervention to strengthen emotional self-regulation in children with emerging mental health problems: proximal impact on school behavior. Journal of Abnormal Child Psychology. 2010; 30:707–720. [PubMed: 20180009]

Yu Q, Tang W, Kowalski J, Tu XM. Multivariate U-statistics: a tutorial with applications. Wiley Interdisciplinary Reviews: Computational Statistics. 2011; 3:457–471.

Yu Q, Chen R, Tang W, He H, Gallop R, Crits-Christoph P, Hu J, Tu XM. Distribution-free models for longitudinal count responses with overdispersion and structural zeros. Statistics inMedicine. 2013; 32(14):2390–2405.

Zhang Z, Wang L. Methods for mediation analysis with missing data. Psychometrika. 2013; 78:154–184. [PubMed: 25107523]

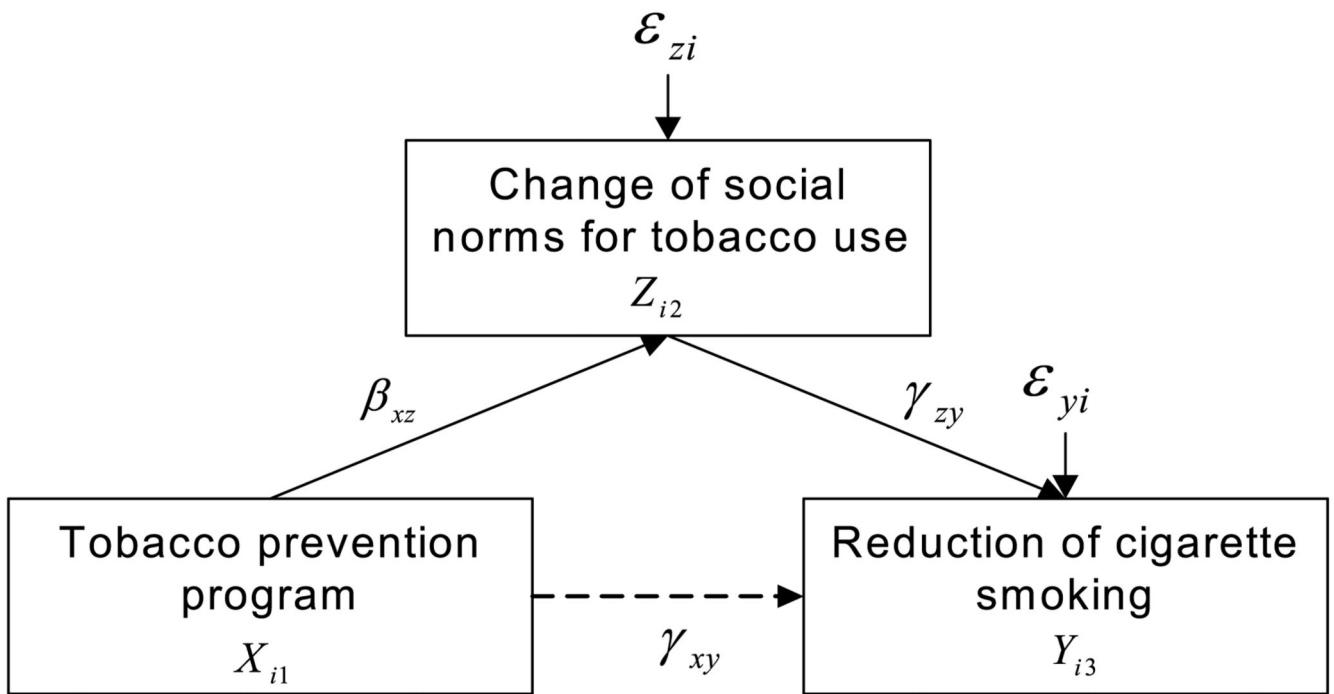**Figure 1.**
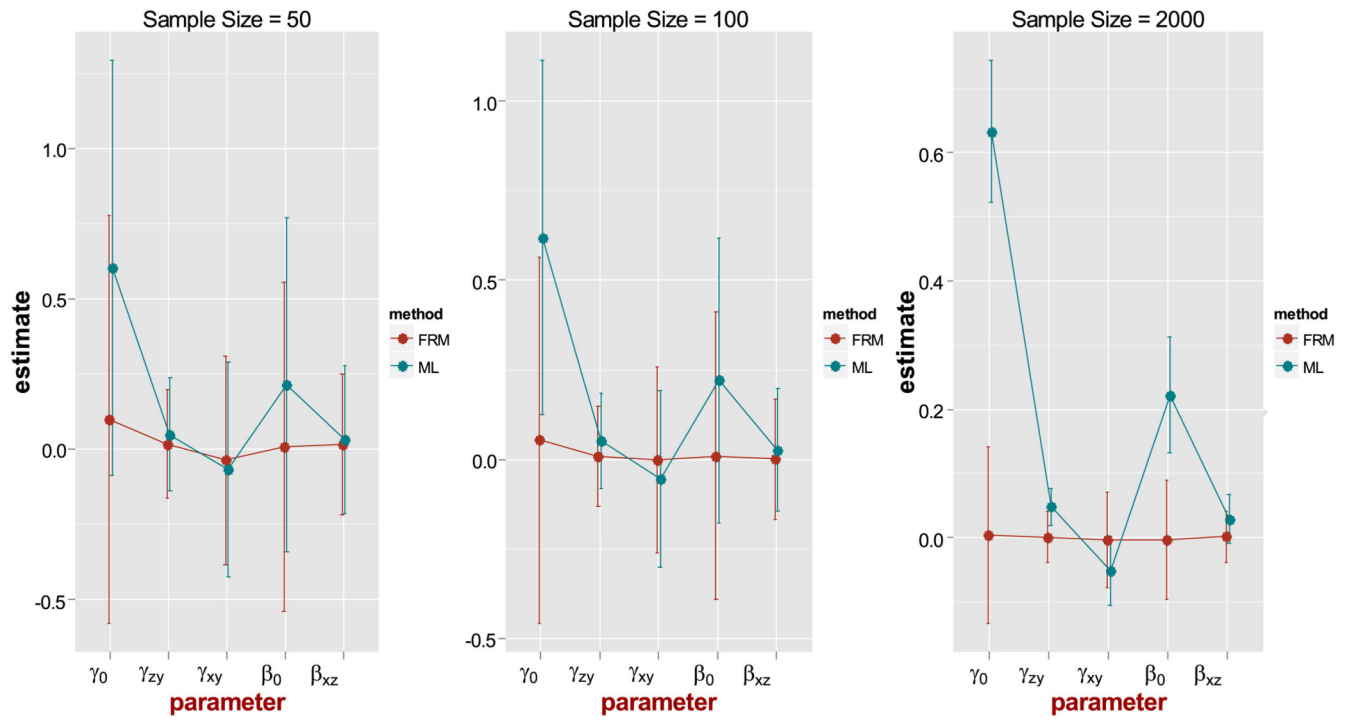Diagram showing the pathway of a mediation process for a tobacco prevention program.

Figure 2.
Simulation Results:Mean estimates – population estimates (± standard errors) show the bias in ML while FRM performs well with missing data.
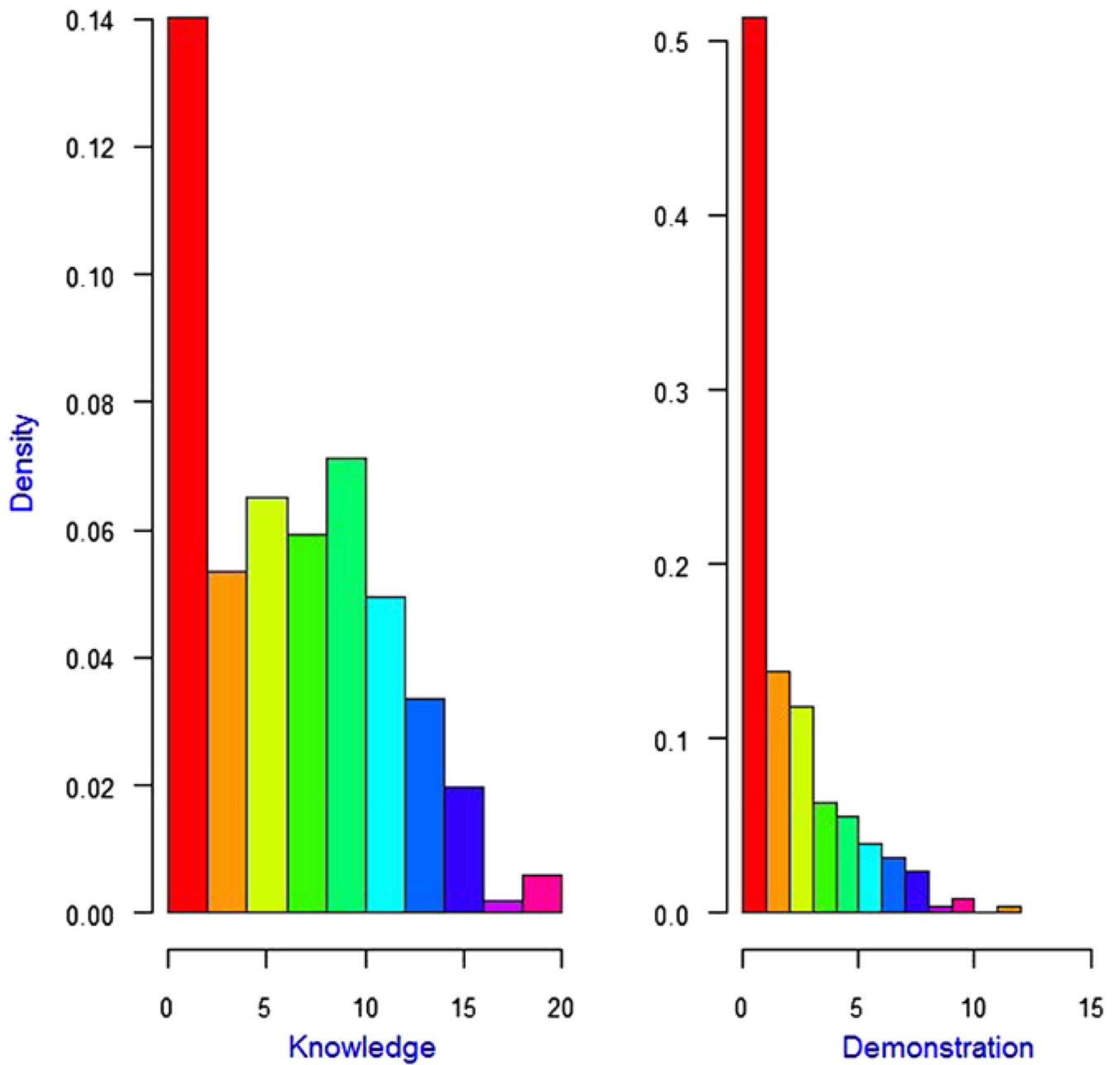
Figure 3.
Histogram of the mediator and outcome for the Child Resilience Study.

Path Diagram from Child Resilience Mediation Analysis

Knowledge of Skills at 6 months

β_{xz}

γ_{zy}

Treatment at Baseline

γ_{xy}

Demonstration of Skills at 6 months
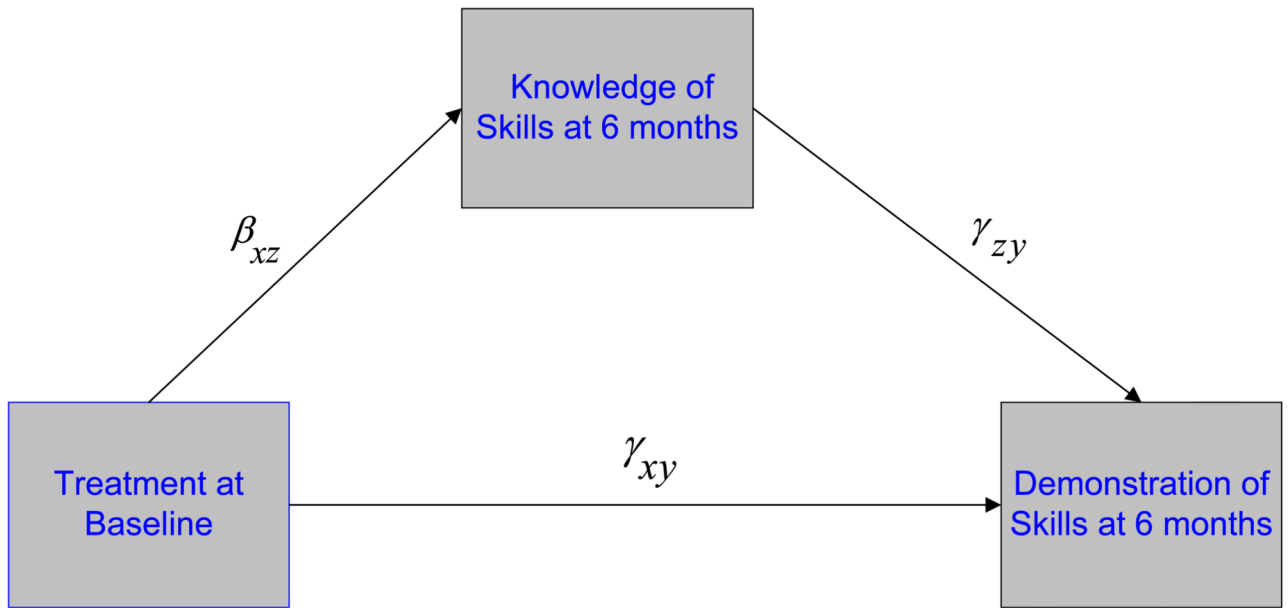
Path Diagram of Direct Effect

1.388
(<.001)

Treatment at Baseline

Demonstration of Skills at 6 months
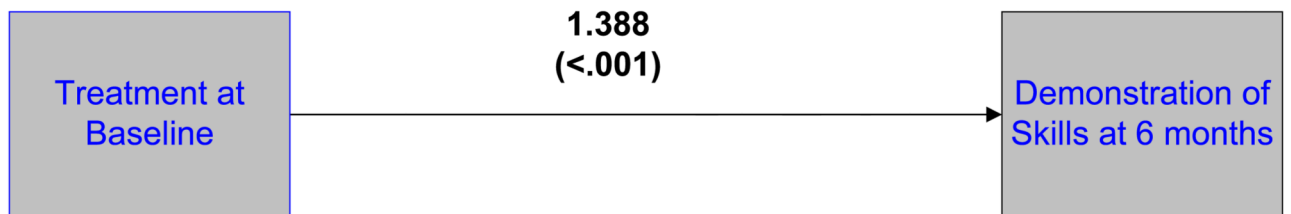
Figure 4.
Path diagram for the mediation model for the Child Resilience Study.

**Path Diagram from Child Resilience
Mediation Analysis with Missing at
Random (MAR) Data**
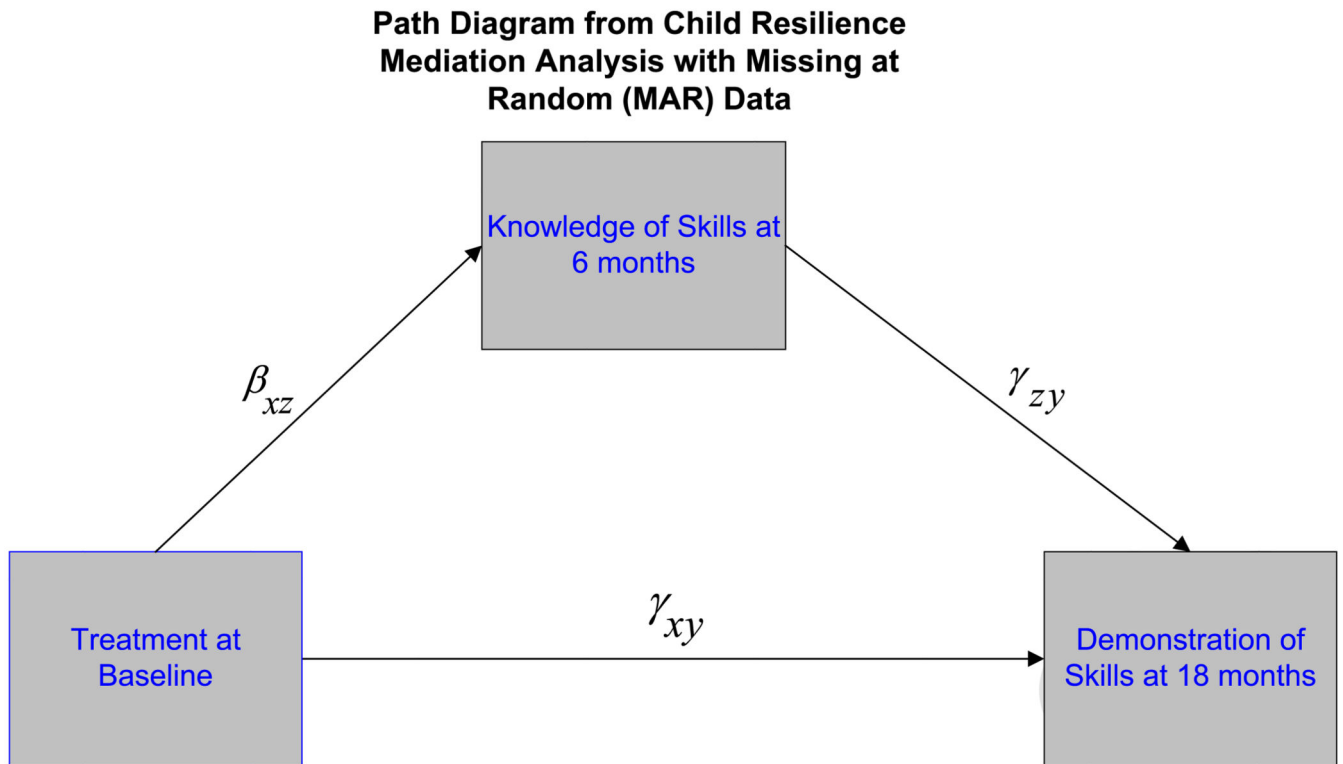
Figure 5.
Path diagram for the mediation model for the Child Resilience Study with MAR Data.

Table 1

Comparison of percent bias of parameter estimates, average asymptotic standard error, relative difference (RD) between average asymptotic and empirical standard error, and type I error rates, under and type I error rates, with missing data under MAR from 1000 MC simulations.

| θ	(%) Bias Method		Standard error Method			
			Asymptotic		RD	
	FRM	ML	FRM	ML	FRM	ML
			Sample size = 50			
γ_0	9.9	60.2	0.679	0.691	0.248	0.107
γ_{zy}	1.7	4.9	0.181	0.187	0.264	0.082
γ_{xy}	−3.7	−6.7	0.347	0.357	0.204	0.062
β_0	0.8	21.4	0.549	0.555	0.070	0.028
β_{xz}	1.4	3.3	0.234	0.246	0.175	0.086
a for H_0: $\gamma_{xy} = 0$			Wald 0.134	0.079		
			Score 0.057			
			Sample size = 100			
γ_0	5.4	61.9	0.512	0.495	0.227	0.078
γ_{zy}	0.9	5.1	0.139	0.133	0.189	0.044
γ_{xy}	−0.1	−5.5	0.259	0.247	0.189	0.004
β_0	1.0	22.1	0.400	0.397	0.046	0.030
β_{xz}	0.2	2.7	0.168	0.171	0.041	0.036
a for H_0: $\gamma_{xy} = 0$			Wald 0.107	0.057		
			Score 0.064			
			Sample size = 2000			
γ_0	0.4	63.3	0.138	0.111	0.148	0.053
γ_{zy}	0.1	4.8	0.040	0.029	0.072	0.034
γ_{xy}	−0.4	−5.1	0.074	0.054	0.115	0.000
β_0	−0.3	22.2	0.093	0.090	0.052	0.033
β_{xz}	0.2	2.9	0.040	0.038	0.000	0.027
a for H_0: $\gamma_{xy} = 0$			Wald 0.054	0.159		
			Score 0.047			

Missing data (15 %/30 % at time 2/3) with correlated repeated measures, t-distributed error terms

Table 2

Comparison of type II error rates through power analysis from 1000 MC simulations.

Power Analysis under H_0: $\gamma_{xy} = 0$, H_A: $\gamma_{xy} = 0.5$					
Normal error under complete data					
	Method				
Sample size	**FRM (Wald)**	**FRM (Score)**	**ML**	**GLS**	**WLS**
50	0.735	0.647	0.706	0.697	0.735
100	0.924	0.899	0.920	0.920	0.924
500	>0.999	>0.999	>0.999	>0.999	>0.999

Missing data (15 %/30 % at time 2/3) with correlated repeated measures, t-distributed error terms			
	Method		
Sample size	**FRM (Wald)**	**FRM (Score)**	**ML**
50	0.350	0.204	0.264
100	0.511	0.363	0.458
500	0.938	0.915	0.985

Table 3

Child Resilience Study complete dataset sample.

ID	Treatment at baseline	Knowledge of skills at 6 months	Demonstration of skills at 6 months
1	No	9	0
2	Yes	8	4
3	No	0	0
4	No	14	3
5	Yes	14	1
6	Yes	0	0
7	No	9	1
8	No	4	1
9	Yes	10	2
.	.	.	.
.	.	.	.

Table 4

Parameter estimates, standard errors, and type I error rates for the mediation model for the Child Resilience Study with complete data.

Estimates, standard errors and type I errors **Child Resilience Study example under complete data**				
θ	**Estimate** **Method**		**Standard error** **Method** **Asymptotic**	
	FRM	**ML**	**FRM**	**ML**
Sample size = 253				
γ_0	0.392	0.392	0.127	0.202
γ_{zy}	0.117	0.117	0.032	0.029
γ_{xy}	0.873	0.873	0.306	0.279
β_0	3.429	3.429	0.369	0.374
β_{xz}	4.390	4.390	0.527	0.529
Type I α for H_0: $\gamma_{xy} = 0$			Wald 0.004	0.002
			Score 0.005	

Table 5

Child Resilience Study dataset sample with missing data.

ID	Treatment at baseline	Knowledge of skills at 6 months	Demonstration of skills at 18 months
1	No	9	3
2	Yes	8	4
3	No	0	NA
4	No	14	0
5	Yes	NA	NA
6	Yes	14	0
7	Yes	NA	NA
8	Yes	0	NA
9	No	9	2
.	.	.	.
.	.	.	.

Table 6

Parameter estimates, standard errors, and p-values for the missing data model for the Child Resilience Study.

Estimates, standard errors and p-value Child Resilience Example under missing data			
η	Estimate	Standard error Asymptotic	p-value
Sample size = 401			
η_{02}	0.546	0.147	<0.001
η_{x1}	−0.019	0.207	0.926
η_{03}	0.250	0.201	0.214
η_{z2}	0.067	0.029	0.022

Table 7

Parameter estimates, standard errors, and type I error rates for the mediation model for the Child Resilience Study with missing data.

θ	Estimates, standard errors and type I errors Child Resilience Study example under missing data (37 %/59 %)			
	Estimate Method		**Standard error Method Asymptotic**	
	FRM	**ML**	**FRM**	**ML**
	Sample size = 401			
γ_0	1.812	1.810	0.278	0.352
γ_{zy}	−0.042	−0.039	0.053	0.050
γ_{xy}	2.330	2.283	0.503	0.480
β_0	3.429	3.429	0.370	0.374
β_{xz}	4.390	4.390	0.528	0.529
Type I α for H_0: $\gamma_{xy} = 0$			Wald < 0.001	<0.001
			Score < 0.001	