Biometrics 65, 104–115
March 2009

DOI: 10.1111/j.1541-0420.2008.01022.x

Residual-Based Diagnostics for Structural Equation Models

B. N. Sánchez,[1,*] **E. A. Houseman,**[2,3] **and L. M. Ryan**[2]

[1]Department of Biostatistics, University of Michigan, School of Public Health, Ann Arbor, Michigan 48104, U.S.A.
[2]Biostatistics Department, Harvard School of Public Health, Boston, Massachusetts 02115, U.S.A.
[3]Department of Work Environment, University of Massachusetts Lowell, Lowell, Massachusetts 01854, U.S.A.
[*]*email:* brisa@umich.edu

Summary. Classical diagnostics for structural equation models are based on aggregate forms of the data and are ill suited for checking distributional or linearity assumptions. We extend recently developed goodness-of-fit tests for correlated data based on subject-specific residuals to structural equation models with latent variables. The proposed tests lend themselves to graphical displays and are designed to detect misspecified distributional or linearity assumptions. To complement graphical displays, test statistics are defined; the null distributions of the test statistics are approximated using computationally efficient simulation techniques. The properties of the proposed tests are examined via simulation studies. We illustrate the methods using data from a study of in utero lead exposure.

Key words: Conditional residuals; Latent-variable residuals; Linearity; Marginal residuals; Normality.

1. Introduction

Structural equation models (SEMs) are becoming increasingly popular in health research (e.g., Bandeen-Roche et al., 1997; Dunson, 2000; Budtz-Joergensen et al., 2002; Liu, Wall, and Hodges, 2005; Proust et al., 2006). This modeling framework is useful in analyzing data from studies where multivariate outcomes or multiple, highly correlated predictors, for example surrogates of an exposure, have been collected. Through the use of latent variables, SEMs succinctly describe associations between multivariate predictors and multivariate outcomes, and alleviate issues of multiple comparisons and collinearity (Sánchez et al., 2005).

Classical fitting methods for these models depend on distributional and linearity assumptions that are important to check. Distributional assumptions are required for using readily available fitting procedures in the presence of data missing at random (Little and Rubin, 2002), for example in the `Mplus` software (Muthén and Muthén, 1998–2004). Although procedures that relax distributional assumptions have been developed (e.g., Browne, 1984; Arminger and Schoenberg, 1989), they work correctly only under data missing completely at random. Further, distributional assumptions are desirable because of improved efficiency, especially when small effect sizes may be expected. Deviations from the assumed distribution lead to biased standard errors, and may hinder the predictive ability of the model. Deviations from linearity are also problematic, but are difficult to explore prior to modeling since the latent variables are unobserved. An incorrect linearity assumption among latent variables leads to, for example, incorrect estimation of exposure effects (e.g., nonlinear dose-response). Further, incorrectly assuming linear relationships between a latent variable and one of its observed indicators is also problematic because it may preclude the comparison of latent variable means across groups (Bauer, 2005). Although procedures that relax linearity assumptions have also been developed (Wall and Amemiya, 2000; Carroll et al., 2004; Lee and Song, 2004), their use has remained relatively limited. Hence, it is of interest to develop tools to check linearity and distributional assumptions for SEMs with latent variables.

Diagnostics tools for SEMs that are based on individual-level residuals is an area that has received little attention, although some notable references exist. Lee and Lu (2003) and Lee and Tang (2004), for example, propose computationally efficient diagnostics to assess the degree of influence a particular observation may have on the estimated model parameters, and model fit. Their methods can be viewed as generalizations of Cook's distance to SEMs. Other authors have used ad hoc diagnostics to check for model fit, such as examining standardized residuals against quantiles from a standard normal distribution. However, such diagnostics do not account for having estimated model parameters with the same data. Not accounting for parameter estimation may affect the Type I error rates of goodness-of-fit tests derived from such ad hoc procedures.

We propose and evaluate a variety of graphical procedures and associated statistical tests to examine distribution and linearity assumptions. Their theoretical basis rests on recent developments in diagnostic methodology for linear mixed models, as follows. Houseman, Ryan, and Coull (2004) and Houseman, Coull, and Ryan (2006) developed theory and methods to test the normality assumption of the error term in a linear model for correlated data, including linear mixed models. Pan and Lin (2005) provide theoretical results and methodology to check the deterministic component of linear mixed models. Both methodologies directly apply to linear SEMs because SEMs have many overlaps with linear mixed models. Specifically, SEMs, like linear mixed models, imply a structured mean and variance for the marginal moments of the data. This similarity is sufficient for the methodologies to apply. We adapt both methods to structural equation modeling. Key issues in adapting their methods to SEMs are defining residuals for the various error components of the model, and adequately weighting observations in the presence of missing data. We evaluate properties of the diagnostics not previously examined because of the complexity of SEMs in contrast to linear mixed models. The proposed methods are graphical in nature, such that they provide information on the type of misspecification. We also provide test statistics to assess significance of lack of fit.

The motivation for this work comes from a study of fetal lead exposure (Tellez-Rojo et al., 2004). Some of the study aims are to quantify the relative contribution of endogenous (e.g., bone lead concentration) versus exogenous (e.g., use of lead contaminated ceramics) sources of exposure and to understand how maternal characteristics impact fetal lead exposure levels. The exposure data collected are high dimensional: various biomarkers and survey data are collected on the mother pre-pregnancy and three or four times during and after pregnancy. Table 1 shows a list of observed biomarkers used in this analysis and their timing. The sample size is 209, although there is a large number of missing data patterns, and higher percentages of missing data occur at the earlier stages of pregnancy when recruitment is more difficult. We model these data with a structural equation model with latent variables, represented in Figure 1, which reduces the dimensionality of the data and succinctly describes associations between lead concentrations in various media.

Table 1
Data collected

Label	Biomarker	Time	N
X_{10}	Plasma	BP	11
X_{20}	Blood ABC Lab	BP	29
X_{30}	Blood Smith Lab	BP	11
X_{11}	Plasma	T1	153
X_{21}	Blood ABC Lab	T1	172
X_{31}	Blood Smith Lab	T1	155
X_{12}	Plasma	T2	169
X_{22}	Blood ABC Lab	T2	173
X_{32}	Blood Smith Lab2	T2	198
X_{13}	Plasma	T3	157
X_{23}	Blood ABC Lab	T3	176
X_{33}	Blood Smith Lab	T3	159
X_{43}	Cord Blood	Birth	107
X_{14}	NTx	T1	95
X_{24}	NTx	T2	127
X_{34}	NTx	T3	137
X_{15}	Patella lead	BP	23
X_{25}	Tibia lead	BP	19
X_{35}	Patella lead	1mpp	203
X_{45}	Tibia lead	1mpp	173
	Covariates		
Z_1	Maternal Age	R	209
Z_2	% of Mother's Life in Mexico City	R	209
Z_3	Frequency of Leaded Ceramics Use	R	209

Notes: BP = Before Pregnancy, 1mpp = 1 month post-partum, Ti = Trimester i, and R = Recruitment.

In Section 2, we define linear structural equation models with latent variables. In Section 3, we propose distribution and linearity diagnostics for SEMs. Section 4 presents results from simulation studies evaluating the empirical size and power of the proposed tests. In Section 5, we demonstrate the utility of these diagnostics by applying them to the study of in utero lead exposure. In Section 6, we state our conclusions and suggest directions for future work.

2. Linear Structural Equation Models

For the ith of n independent units, let $\mathbf{X}_i$ represent p error-prone measurements of an ℓ-dimensional latent variable vector $\mathbf{U}_i$, and let $\mathbf{Z}_i$ represent q fixed covariates. We specify a two-stage linear SEM for this data and name the sources of error in the model. We also write marginal moments for the observed data, which are later used in defining diagnostics.

The first stage relates dependent variables, $\mathbf{X}_i$, to latent variables and fixed covariates

$$\mathbf{X}_i = \boldsymbol{\nu} + \boldsymbol{\Lambda}\mathbf{U}_i + \mathbf{K}\mathbf{Z}_i + \boldsymbol{\epsilon}_i, \tag{1}$$

where $\boldsymbol{\nu}_{p\times 1}$, $\boldsymbol{\Lambda}_{p\times \ell}$, and $\mathbf{K}_{p\times q}$ are parameter matrices with some elements often restricted to zero or one to ensure identifiability (Sánchez et al., 2005). The vector of *conditional errors* $\boldsymbol{\epsilon}_i$ is multivariate normal with $E(\boldsymbol{\epsilon}_i \mid \mathbf{U}_i, \mathbf{Z}_i) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}_i \mid \mathbf{U}_i, \mathbf{Z}_i) = \boldsymbol{\Sigma}_\epsilon$. This error term captures the deviation of the dependent variables from their conditional mean, given the latent variables and covariates, for example, measurement error. Measurement errors within a subject may be correlated (see Figure 1) thus $\boldsymbol{\Sigma}_\epsilon$ may not always be diagonal.

The second stage of the model defines linear relationships between the latent variables

$$\mathbf{U}_i = \boldsymbol{\alpha} + \mathbf{B}\mathbf{U}_i + \boldsymbol{\Gamma}\mathbf{Z}_i + \boldsymbol{\zeta}_i, \tag{2}$$

where $\boldsymbol{\Gamma}_{\ell\times q}$ and $\mathbf{B}_{\ell\times\ell} = \{\beta_{gh}\}$ ($\beta_{gg} = 0$ for all g) may also have entries restricted to zero. The vector of *latent variable errors* $\boldsymbol{\zeta}_i$ is independent of $\boldsymbol{\epsilon}_i$, and is normally distributed, with $E(\boldsymbol{\zeta}_i \mid \mathbf{Z}_i) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\zeta}_i \mid \mathbf{Z}_i) = \boldsymbol{\Psi}$. In the lead study, latent circulating lead level is a longitudinal measure, thus elements of $\boldsymbol{\zeta}_i$ may be correlated, leading to nondiagonal $\boldsymbol{\Psi}$.

Let $\boldsymbol{\theta}$ represent all free parameters that parameterize $\boldsymbol{\nu}, \boldsymbol{\Lambda}, \mathbf{K}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}_\epsilon, \boldsymbol{\alpha}, \boldsymbol{\Gamma}$, and $\boldsymbol{\Psi}$. As they will later be useful in defining residuals, we state the following marginal moments

$$\begin{aligned}\boldsymbol{\mu}_i \equiv \boldsymbol{\mu}(\boldsymbol{\theta}; \mathbf{Z}_i) = E\left(\mathbf{X}_i \mid \mathbf{Z}_i\right) &= \boldsymbol{\nu} + \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\alpha} \\ &\quad + [\boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma} + \mathbf{K}]\mathbf{Z}_i \end{aligned} \tag{3}$$

$$\boldsymbol{\Sigma} \equiv \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathrm{Var}\left(\mathbf{X}_i \mid \mathbf{Z}_i\right) = \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Psi}(\mathbf{I} - \mathbf{B})^{-T}\boldsymbol{\Lambda}^T + \boldsymbol{\Sigma}_\epsilon. \tag{4}$$

Note that unlike linear mixed models, where it is often assumed that the marginal mean and variance are parameterized by two distinct sets of parameters (Longford, 1993, p. 26), the marginal mean (3) and variance (4) share parameters ($\boldsymbol{\Lambda}$ and $\mathbf{B}$). We use maximum likelihood estimation to obtain parameter estimates, but refer the reader elsewhere for estimation details (e.g., Bollen, 1989; Sánchez et al., 2005).

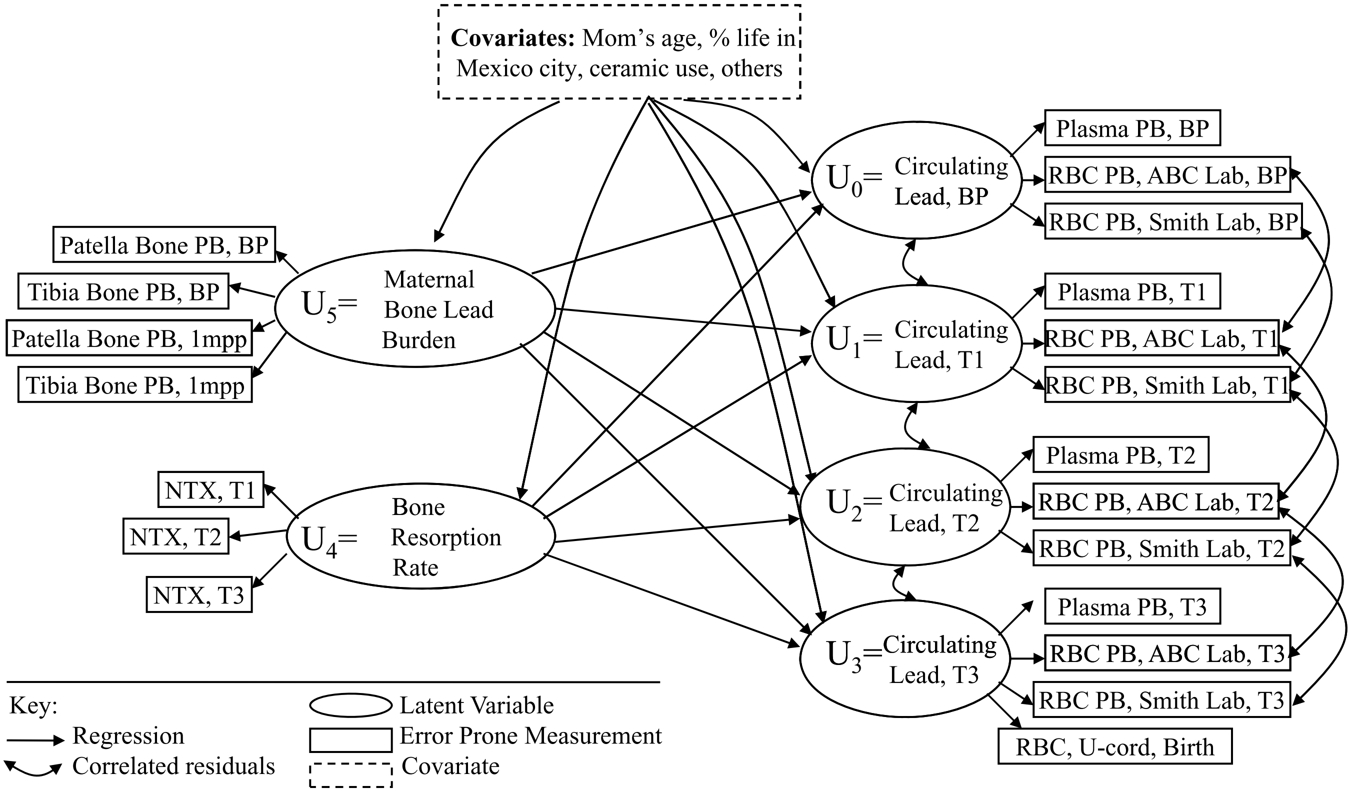

Figure 1. Model describing association between latent bone lead concentration, latent bone resorption rate, and latent circulating lead concentrations. Latent variables are indirectly measured by various biomarkers, and are regressed on fixed covariates. Circulating lead levels depend additionally on bone resorption rate and bone lead concentration. Abbreviations: See Table 1 legend, and PB = Lead, NTX = Bone resorption biomarker, RBC = Red Blood Cell.

3. Residual-Based Diagnostics

The model specification assumes linearity, and maximum likelihood estimation requires distributional assumptions; either or both assumptions could be violated. The distributions of either the latent variable errors, or the conditional errors, or both, can deviate from normality. Given the linearity assumptions at both stages of the model, there are at least four possibilities for misspecification. In (2), one could misspecify (a) the functional form of a covariate on a latent variable or (b) the link between a latent variable and its linear predictor; and in (1), (c) the functional form of a covariate on a surrogate or (d) the link between a surrogate and its linear predictor. To propose diagnostics to check these assumptions, we first define several types of residuals. Residuals similar to ours have been described previously by Bollen and Arminger (1991).

Standardized marginal residuals for the ith subject are defined as

$$\mathbf{r}_i = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X}_i - \boldsymbol{\mu}_i), \qquad (5)$$

where $\boldsymbol{\Sigma}^{1/2}$ is the square root of $\boldsymbol{\Sigma}$. Because these residuals have zero mean, variance one, and are uncorrelated, they lend themselves to regression diagnostics for independent data. Because the marginal residuals encompass both the conditional and latent variable error, they will serve to define an omnibus test of fit.

Next we define the kth conditional and gth latent variable residuals based on standardized estimates of the respective components of the conditional and latent variable error vectors. Estimates of the errors ϵ_{ik} and ζ_{ig} are obtained by taking their conditional expectation, given the observed data. As detailed in Web Appendix A, $E(\epsilon_{ik} \mid \mathbf{X}_i) = \boldsymbol{\pi}_k^c \boldsymbol{\Sigma}_\epsilon \boldsymbol{\Sigma}^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_i)$, where $\boldsymbol{\pi}_k^c$ is a $(1 \times p)$ matrix that selects the kth component of $E(\boldsymbol{\epsilon}_i \mid \mathbf{X}_i)$, i.e., it is zero everywhere except at the (1, k) position, which is one. Similarly, $E(\zeta_{ig} \mid \mathbf{X}_i) = \boldsymbol{\pi}_g^\ell \boldsymbol{\Psi}(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_i)$, where $\boldsymbol{\pi}_g^\ell$ selects the gth component of $E(\boldsymbol{\zeta}_i \mid \mathbf{X}_i)$. To standardize, first note that the conditional expectations above can be written as $E(\epsilon_{ik} \mid \mathbf{X}_i) = \mathbf{C}_k\boldsymbol{\Sigma}^{-1/2}(\mathbf{X}_i - \boldsymbol{\mu}_i)$ and $E(\zeta_{ig} \mid \mathbf{X}_i) = \mathbf{D}_g\boldsymbol{\Sigma}^{-1/2}(\mathbf{X}_i - \boldsymbol{\mu}_i)$, where $\mathbf{C}_k = \boldsymbol{\pi}_k^c\boldsymbol{\Sigma}_\epsilon\boldsymbol{\Sigma}^{-1/2}$ and $\mathbf{D}_g = \boldsymbol{\pi}_g^\ell\boldsymbol{\Psi}(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1/2}$. Then, letting $\mathbf{P}_k^c = (\mathbf{C}_k\mathbf{C}_k^T)^{-1/2}\mathbf{C}_k$, and $\mathbf{P}_g^\ell = (\mathbf{D}_g\mathbf{D}_g^T)^{-1/2}\mathbf{D}_g$, we arrive at

$$r_i^{c,k} = \mathbf{P}_k^c\mathbf{r}_i, \text{ (conditional)} \qquad (6)$$

$$r_i^{\ell,g} = \mathbf{P}_g^\ell\mathbf{r}_i \text{ (latent)}, \qquad (7)$$

which are standardized conditional and latent variable residuals. Standardizing is important in the presence of missing data, as the estimated residuals will have different variabilities. Web Appendix B details how the definitions of the residuals are affected by missing data. Also note that the conditional and latent variable residuals are linear combinations, or projections, of the marginal residuals $\mathbf{r}_i$ defined in (5). The projections play a role in defining the diagnostics.

3.1 *Distribution Diagnostics*

To assess normality assumptions, we first describe the empirical cumulative distribution functions (ECDFs) of the above defined residuals, and later define ECDF-based test statistics. Because the asymptotic distribution of the test statistics is difficult to obtain analytically, we then discuss an efficient simulation technique to compute p-values.

To construct an omnibus test of normality, consider the ECDF of the standardized marginal residuals $\mathbf{r} = (\mathbf{r}_1^T, \mathbf{r}_2^T, \dots, \mathbf{r}_n^T)^T$. For a fixed value of t in the real line, this ECDF is

$$F_{pn}(t; \boldsymbol{\theta}, \mathbf{r}) = (pn)^{-1} \sum_{i=1}^{n} \sum_{k=1}^{p} 1\{r_{ik} \le t\}, \qquad (8)$$

where r_{ik} is the kth component of the ith subject's standardized marginal residuals, $\mathbf{r}_i$. Under the null hypothesis that $\boldsymbol{\epsilon}_i$ and $\boldsymbol{\zeta}_i$ are normally distributed, and assuming that the marginal mean, $\boldsymbol{\mu}_i$, and variance, $\boldsymbol{\Sigma}$, are correctly specified, then (8) evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, the true value of $\boldsymbol{\theta}$, approximates the standard normal cumulative distribution function, $\Phi(t)$. Large deviations of (8) from $\Phi(t)$ would indicate that either the latent variable errors, or the conditional errors, or both, are not normally distributed. To construct tests for lack of normality for the kth or gth conditional errors latent variable errors, ϵ_{ik} or ζ_{ig}, consider the ECDF of the residuals in (6) and (7). For a given matrix $\mathbf{P}$ (e.g., $\mathbf{P} = \mathbf{P}_k^c$ or $\mathbf{P} = \mathbf{P}_g^\ell$), the ECDF of the residuals is

$$F_n(t; \boldsymbol{\theta}, \mathbf{P}, \mathbf{r}) = n^{-1} \sum_{i=1}^{n} 1\{\mathbf{P}\mathbf{r}_i \le t\}. \qquad (9)$$

Thus, for example, under the null hypothesis that ϵ_k is normally distributed, and assuming that $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}$ are correctly specified, $F_n(t;\, \boldsymbol{\theta} = \boldsymbol{\theta}_0,\, \mathbf{P} = \mathbf{P}_k^c,\, \mathbf{r})$ should approximate $\Phi(t)$.

Standard normality tests (van der Vaart, 1998) cannot be conducted to assess the significance of the observed differences between the ECDFs and $\Phi(t)$ because $\boldsymbol{\theta}_0$ is unknown. Instead, a variety of test statistics capturing the discrepancy between the ECDFs and $\Phi(t)$ can be defined. For instance, we compute test statistics similar to the Kolmogorov–Smirnov (KS) test and the Cramér–Von Mises (CVM) test, respectively,

$$\tau_{\mathrm{KS}} = \sup \left| F_n(t; \widehat{\boldsymbol{\theta}}, \widehat{\mathbf{P}}, \mathbf{r}) - \Phi(t) \right|, \quad \text{and}$$
$$\tau_{\mathrm{CVM}} = \int (F_n(t; \widehat{\boldsymbol{\theta}}, \widehat{\mathbf{P}}, \mathbf{r}) - \Phi(t))^2 d\Phi(t).$$

The null distribution of test statistics like τ_{KS} and τ_{CVM}, which adjust for the estimation of $\widehat{\boldsymbol{\theta}}$, are difficult to obtain analytically but can be approximated by simulation (Houseman et al., 2004, 2006). That is, first simulate the behavior of the discrepancies $F_{pn}(t; \widehat{\boldsymbol{\theta}}, \mathbf{r}) - \Phi(t)$ or $F_n(t; \widehat{\boldsymbol{\theta}}, \widehat{\mathbf{P}}, \mathbf{r}) - \Phi(t)$ under the null hypothesis; then compute the statistics, τ_{KS} and τ_{CVM}, over many simulations; and finally obtain a p-value by calculating the proportion of simulated test statistics that are larger than the observed statistic.

Under the null hypothesis, the behavior of $F_{pn}(t; \widehat{\boldsymbol{\theta}}, \mathbf{r}) - \Phi(t)$ and $F_n(t; \widehat{\boldsymbol{\theta}}, \widehat{\mathbf{P}}, \mathbf{r}) - \Phi(t)$ can be simulated by computing realizations of the stochastic processes in t

$$\begin{aligned}\widehat{F}^*_{pn}(t) - \Phi(t) &= F_{pn}(t; \widehat{\boldsymbol{\theta}}, \mathbf{r}^*) - \Phi(t)\\ &\quad + \delta_{pn}(t, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}})^T \mathbf{J}(\widehat{\boldsymbol{\theta}})^{-1} \mathbf{S}(\widehat{\boldsymbol{\theta}}, \mathbf{r}^*), \quad \text{and} \quad (10)\end{aligned}$$

$$\begin{aligned}\widehat{F}^*_{n}(t) - \Phi(t) &= F_{n}(t; \widehat{\boldsymbol{\theta}}, \mathbf{P}, \mathbf{r}^*) - \Phi(t)\\ &\quad + \delta_{n}(t, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}})^T \mathbf{J}(\widehat{\boldsymbol{\theta}})^{-1} \mathbf{S}(\widehat{\boldsymbol{\theta}}, \mathbf{r}^*). \qquad (11)\end{aligned}$$

As detailed in Web Appendix C, (10)–(11) are derived from first-order Taylor series expansions of the corresponding ECDFs around $\boldsymbol{\theta}_0$, evaluated at $\boldsymbol{\theta}_0 = \widehat{\boldsymbol{\theta}}$. In (10) and (11), $\mathbf{P} = \mathbf{P}(\boldsymbol{\theta})$; $\mathbf{r}^* = (\mathbf{r}_1^{*T}, \mathbf{r}_2^{*T}, \dots \mathbf{r}_n^{*T})^T$, where $\mathbf{r}_i^*$ are $p \times 1$ vectors of simulated standard normal deviates; $\mathbf{J}(\widehat{\boldsymbol{\theta}})$ is the expected information matrix evaluated at $\widehat{\boldsymbol{\theta}}$; and $\mathbf{S}(\widehat{\boldsymbol{\theta}}, \mathbf{r}^*)$ is the score function for $\boldsymbol{\theta}$, with the observed data vector for the ith subject, $\mathbf{X}_i$, replaced by $\boldsymbol{\Sigma}^{-T}\mathbf{r}_i^* + \boldsymbol{\mu}(\boldsymbol{\theta}, \mathbf{Z}_i)$. Finally, $\delta_{pn}(t, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}})$ and $\delta_n(t, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}})$ are, respectively, the $\boldsymbol{\theta}$ derivatives of the expected value of the corresponding ECDF, evaluated at $(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = (\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}})$. Expressions for $\delta_{pn}(t, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}})$ and $\delta_n(t, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}})$ are also given in Web Appendix C.

3.2 *Diagnostics for Linearity*

To assess departures from linearity, we propose tests based on cumulative sums of projected residuals, $r_i^{c,k}$ or $r_i^{\ell,g}$, taken with respect to covariate values or certain predicted values (e.g., $E(X_{ik})$ or $E(U_{ig})$). These sums are stochastic processes which, under the null hypothesis, fluctuate about zero (Pan and Lin, 2005). In contrast to residual plots, the asymptotic behavior of the sums can be approximated via simulation such that inference, and thus objective conclusions, regarding lack of linearity can be drawn. Further, the shape of the cumulative sum is informative about the type of nonlinearity as explained in the example in Section 5, and further detailed by Lin, Wei, and Ying (2002).

Misspecified covariate effects on the gth latent variable in (2) can be assessed by summing the latent variable residuals, $r_i^{\ell,g}$, with respect to the jth covariate

$$W^\ell_{Z_j}(t) = n^{-1/2} \sum_{i=1}^{n} 1\{Z_{ij} \le t\} r_i^{\ell,g}. \qquad (12)$$

Tests can also be defined for testing the link function between the gth latent variable and its linear predictor by considering the cumulative sum of latent variable residuals

$$W_{U_g}(t) = n^{-1/2} \sum_{i=1}^{n} 1\{E(U_{ig}) \le t\} r_i^{\ell,g}, \qquad (13)$$

where $E(U_{ig}) = \boldsymbol{\pi}_g^\ell E(\mathbf{U}_i) = \boldsymbol{\pi}_g^\ell\ (\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\alpha} + \boldsymbol{\Gamma}\mathbf{Z}_i)$.

Misspecified relationships between a surrogate and one of its predictors (e.g., Z_j or U_g) can be assessed by summing the kth conditional residuals with respect to Z_j or $E(U_g)$

$$\begin{aligned}W^c_{Z_j}(t) &= n^{-1/2} \sum_{i=1}^{n} 1\{Z_{ij} \le t\} r_i^{c,k} \quad \text{or}\\ W^c_{U_g}(t) &= n^{-1/2} \sum_{i=1}^{n} 1\{E(U_{ig}) \le t\} r_i^{c,k}.\end{aligned} \qquad (14)$$

The sum $W^c_{Z_j}(t)$ can be used to diagnose *item bias*, that is, a subject's differential response to a surrogate (item) due to a covariate, conditional on the latent variables (Beck, 1982). Finally, similar to testing the link between a latent variable and its linear predictor,

$$W_{X_k}(t) = n^{-1/2} \sum_{i=1}^{n} 1\{E(X_{ik}) \le t\} r_i^{c,k}, \tag{15}$$

can be used to assess the link between a surrogate and its linear predictor.

To draw inference, it is necessary to study the behavior of the cumulative sums (12)–(15) under the null hypothesis of correctly specified (linear) associations. The cumulative sums of residuals define stochastic processes in t, which, under the null hypothesis, fluctuate around zero (Pan and Lin, 2005). Hence, unusually large departures from zero are indicative of model misspecification. Departures from zero can be characterized by test statistics defined as functionals, $\Im(\cdot)$, of the cumulative sums; for example, the sup and L_2 norms,

$$\Im_\infty(W(t)) = \sup_t |\, W(t)\,|, \quad \text{and} \quad \Im_2(W(t)) = \int_t (W(t))^2 \, dt,$$

can be used. These are analogous to the KS and CVM tests for normality.

The asymptotic behavior of such test statistics, however, is difficult to obtain analytically. Instead, the null distribution of the functional-based test statistics, $\Im(\cdot)$ is approximated via simulation. Specifically, we draw realizations of the stochastic processes (12)–(15) under the null hypothesis; calculate the statistics on each of many realizations; and compute p-values as the proportion of simulated statistics that exceed the observed test statistic. Under the null hypothesis, realizations of the cumulative sums of residuals can be simulated as follows. First, note that (12) and (14) are special cases of a multivariate stochastic process

$$W_Z(\mathbf{t}) = n^{-1/2} \sum_{i=1}^{n} 1\{\mathbf{Z}_i \le \mathbf{t}\} \mathbf{P} \mathbf{r}_i, \tag{16}$$

where $\mathbf{t} = (t_1, t_2, \ldots, t_q)$, and $1\{\mathbf{Z}_i \le \mathbf{t}\} = \prod_{k=1}^{p} 1\{Z_k \le t_k\}$, and $\mathbf{P}$ is $\mathbf{P}^c_k$ or $\mathbf{P}^\ell_g$. The cumulative sum with respect to one covariate Z_j is obtained from (16) by setting $t_k = \infty$ for all $k \ne j$. Under the null hypothesis, (16) can be shown (Pan and Lin, 2005) to converge in distribution to the conditional distribution, given the data $\mathbf{Z}_i$, $\mathbf{X}_i$, of the following zero-mean Gaussian process

$$\widehat{W}_Z(\mathbf{t}) = n^{-1/2} \sum_{i=1}^{n} \big(1\{\mathbf{Z}_i \le \mathbf{t}\} \mathbf{P} \mathbf{r}_i + \vartheta(\widehat{\boldsymbol{\theta}}, \mathbf{t})^T \mathbf{J}^{-1}(\widehat{\boldsymbol{\theta}}) \mathbf{S}_i(\widehat{\boldsymbol{\theta}}; \mathbf{X}_i)\big) G_i, \tag{17}$$

where $(G_1,\, G_2, \ldots\, G_n)$ are *i.i.d.* $\sim N(0,\, 1)$, independent of the observed data, and $\vartheta(\boldsymbol{\theta}, \mathbf{t}) = -n^{-1} \sum_{i=1}^{n} (1\{\mathbf{Z}_i \le \mathbf{t}\} \frac{\partial \mathbf{P} \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}})$. Repeated realizations of $\widehat{W}_Z(\mathbf{t})$ can be obtained by fixing the observed data, and drawing samples of $(G_1,\, G_2,\, \ldots\, G_n)$. The approximation (17) of $W_Z(\mathbf{t})$ is derived from a Taylor series expansion. The theory behind (17) is different from that for the distribution diagnostics. The simulation technique for the distribution tests, τ_{KS} and τ_{CMV}, is an approximate parametric bootstrap (Houseman et al., 2006), which requires re-sampling from the null distribution of marginal residuals, $\mathbf{r}^*$ (i.e., pn samples), and evaluates $\mathbf{S}(\widehat{\boldsymbol{\theta}}, \mathbf{r}^*)$ for each sample $\mathbf{r}^*$. In contrast (17) is based on a conditional multiplier central limit theorem (Su and Wei, 1991; Pan and Lin, 2005), requires n samples, $G's$, and evaluates $\mathbf{S}_i(\widehat{\boldsymbol{\theta}}, \mathbf{X}_i)$ once.

The processes $\widehat{W}_{U_g}(t)$ and $\widehat{W}_{X_k}(t)$ can be defined similarly to (17) with $1\{\mathbf{Z}_i \le \mathbf{t}\}$ replaced by $1\{E(U_{ig}) \le t\}$ and $1\{E(X_{ik}) \le t\}$, respectively. Proof of the weak convergence of $W_Z(t), \widehat{W}_Z(t),\ W_{U_g}(t), \widehat{W}_{U_g}(t),\ W_{X_k}(t)$, and $\widehat{W}_{X_k}(t)$ follows from the convergence of the *weighted* processes considered in Section 3.5 of Pan and Lin (2005).

4. Simulation Studies

We conducted simulation studies to evaluate the performance of the proposed diagnostics. Although simulation studies evaluating the properties of these methods have been conducted in the context of linear mixed models (Pan and Lin, 2005; Houseman et al., 2006), there is a need to examine the performance of these methods in assessing model misspecifications unique to structural equation modeling.

To evaluate the empirical size and power of the tests, the model: $\mathbf{X}_i = \boldsymbol{\nu} + \boldsymbol{\Lambda} \mathbf{U}_i + \boldsymbol{\epsilon}_i$ and $\mathbf{U}_i = \boldsymbol{\alpha} + \mathbf{B} \mathbf{U}_i + \boldsymbol{\Gamma} \mathbf{Z}_i + \boldsymbol{\zeta}_i$, with $p = 6, \ell = 2, q = 2$ was fitted to simulated data. The form of the parameter matrices was always assumed to be $\boldsymbol{\nu} = (\nu_1,\, \nu_2, \ldots,\, \nu_6)^T, \mathbf{K} = \mathbf{0}, \boldsymbol{\alpha} = (0,\, 0)$,

$$\boldsymbol{\Lambda}^T = \begin{pmatrix} 1 & \lambda_2 & \lambda_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \lambda_5 & \lambda_6 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 0 & 0 \\ \beta & 0 \end{pmatrix}, \quad \text{and}$$

$$\boldsymbol{\Gamma} = \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix}.$$

To estimate the tests' empirical sizes, we generated 2000 data sets from the assumed model, and used 1000 simulated test statistics to calculate empirical rejection probabilities. Parameter values were set to $\boldsymbol{\nu} = (0,\, 1,\, 2,\, 0,\, 1,\, 2), \mathbf{K} = \mathbf{0}, \boldsymbol{\alpha} = (0,\, 0), \lambda_2 = 0.5, \lambda_3 = 1.25, \lambda_5 = 0.5, \lambda_6 = 1.25, \beta = 1, \gamma_{11} = 1, \gamma_{12} = 0.5,\ \gamma_{21} = 1$, and $\gamma_{22} = 1$. The conditional errors were assumed uncorrelated, normally distributed, and with variance parameters chosen such that the measurement error variance for all surrogates accounted for 30% or 50% of their marginal variance. The latent variable residuals, ζ_1 and ζ_2, were assumed to be uncorrelated and have unit variance. Covariates, $\mathbf{Z}_i$, were simulated from a bivariate normal distribution with mean (0,0), unit variance, and correlation 0.2. To evaluate power, we generated 1000 data sets from each alternate model, described below, and used 1000 realizations from the distributions of the test statistics under the null hypothesis to calculate empirical rejection probabilities. All model parameters were estimated in `Mplus` (Muthén and Muthén, 1998–2004), and the linearity and normality tests were implemented in the statistical package `R` (functions available at `http://www.biometrics.tibs.org`).

4.1 *Distribution Diagnostics*

We evaluated the power of the distribution tests to detect lack of normality in a specific component of the latent variable error or in a specific component of the conditional error. We simulated data using skewed (centered χ^2_3) or heavy-tailed (t_3) distributions. We considered scenarios where either the conditional error ϵ_{i1} or the latent variable error ζ_{i1} was nonnormal,

Table 2

Empirical size and power for distribution diagnostics, tests of nominal size 0.05

CVM		Surrogates with 30% meas. error			Surrogates with 50% meas. error		
alternative	n	Test ϵ_1	Test ζ_1	Omnibus	Test ϵ_1	Test ζ_1	Omnibus
Null Model	200	0.048	0.049	0.053	0.049	0.05	0.048
($\boldsymbol{\zeta}, \boldsymbol{\epsilon} \sim$ Normal)	300	0.046	0.048	0.058	0.048	0.061	0.052
(1) ζ_1 Skewed	200	0.051	0.99	0.11	0.057	0.76	0.07
ϵ_1 Normal	300	0.045	>0.99	0.17	0.053	0.92	0.084
(2) ζ_1 Normal	200	0.98	0.059	0.28	>0.99	0.061	0.4
ϵ_1 Skewed	300	>0.99	0.053	0.38	>0.99	0.063	0.6
(3) ζ_1 Skewed	200	>0.99	>0.99	0.62	>0.99	0.83	0.64
ϵ_1 Skewed	300	>0.99	>0.99	0.83	>0.99	0.96	0.84
(4) ζ_1 Heavy Tailed	200	>0.99	0.75	0.47	>0.99	0.52	0.53
ϵ_1 Skewed	300	>0.99	0.88	0.65	>0.99	0.62	0.72
(5) ζ_1 Skewed	200	0.76	>0.99	0.37	0.84	0.79	0.38
ϵ_1 Heavy Tailed	300	0.87	>0.99	0.52	0.94	0.94	0.52

KS		Surrogates with 30% meas. error			Surrogates with 50% meas. error		
alternative	n	Test ϵ_1	Test ζ_1	Omnibus	Test ϵ_1	Test ζ_1	Omnibus
Null Model	200	0.055	0.053	0.055	0.057	0.053	0.051
($\boldsymbol{\zeta}, \boldsymbol{\epsilon} \sim$ Normal)	300	0.044	0.040	0.046	0.039	0.061	0.052
(1) ζ_1 Skewed	200	0.053	0.63	0.057	0.054	0.65	0.57
ϵ_1 Normal	300	0.054	>0.99	0.13	0.048	0.85	0.09
(2) ζ_1 Normal	200	0.98	0.055	0.3	0.98	0.067	0.3
ϵ_1 Skewed	300	>0.99	0.05	0.28	>0.99	0.052	0.47
(3) ζ_1 Skewed	200	0.98	0.7	0.49	>0.99	0.72	0.51
ϵ_1 Skewed	300	>0.99	>0.99	0.69	>0.99	0.9	0.7
(4) ζ_1 Heavy Tailed	200	0.99	0.39	0.4	>0.99	0.41	0.41
ϵ_1 Skewed	300	0.99	0.8	0.48	>0.99	0.52	0.57
(5) ζ_1 Skewed	200	0.74	0.7	0.26	0.74	0.71	0.26
ϵ_1 Heavy Tailed	300	0.78	>0.99	0.38	0.87	0.87	0.39

as well as cases where both ϵ_{i1} and ζ_{i1} deviated from normality. We conducted tests of normality on the conditional and the latent variable residuals, as well as the omnibus test.

Table 2 illustrates the empirical size and power for tests of nominal size 0.05. The top half of the table shows results for the CVM tests, while the bottom half shows results for the KS tests. In general, the simulation-based tests proposed here perform well in terms of power, and are within Monte Carlo error of the nominal size. For every type of misspecification, the CVM test was more powerful than the KS test. The test for lack of normality of the conditional error was, in general, more powerful than the test for the latent variable error (e.g., compare Alternatives 1 and 2 under the 50% error column). Similarly, the omnibus test had slightly more power to detect a misspecified distribution for ϵ_{i1} than ζ_{i1}, and had high power when both distributions were misspecified. As expected, larger measurement error variance decreases the power of the tests. Finally, and most importantly, the tests were specific to the stage of the model containing the deviation from normality. If the conditional error was not normal while the latent variable residual was normal, then the test conducted on the latent variable residuals preserved its Type-1 error rate, and vice versa.

4.2 *Linearity Diagnostics*

We evaluated the performance of linearity tests based on cumulative residuals. In scenarios where the functional form of a covariate is misspecified, at either stage of the model, we also evaluated the power of the corresponding link test, and the CVM normality test (Eberly and Thackeray, 2005, show that normality tests can detect misspecified mean functions). For comparison, we also evaluate the power of the Wald test of the null hypothesis when fitting the model under the (correct) alternative hypothesis. The Wald test was only possible for cases when the null was nested in the alternative.

To investigate the power of the test of the functional form of a covariate for the mean of a latent variable, we considered three alternatives for the structural part of the model

$$\begin{aligned}
\textbf{Model 1a:}\quad U_{1i} &= \gamma_{11} Z_{1i} + \gamma_{12} Z_{2i} + \gamma_{13} Z_{2i}^2 + \zeta_{1i},\\
\textbf{Model 1b:}\quad U_{1i} &= \gamma_{11} Z_{1i} + \gamma_{12} \exp(Z_{2i}) + \zeta_{1i},\\
\textbf{Model 1c:}\quad U_{1i} &= \gamma_{11} Z_{1i} + \gamma_{12} Z_{2i} + \gamma_{13} Z_{3i} + \zeta_{1i},
\end{aligned}$$

where $Z_{3i} \sim$ Normal(0,1), and independent from (Z_{1i}, Z_{2i}). Models 1a and 1b represent scenarios where a quadratic term should be included in the linear predictor for U_{1i}, and where the covariate was incorrectly log-transformed. Model 1c implies Z_{3i} is missing from the model specification. We use data on Z_3 to construct the cumulative sum diagnostic; for example, mother's age in the example section, Figure 2c. The power of the link test between covariates and a latent

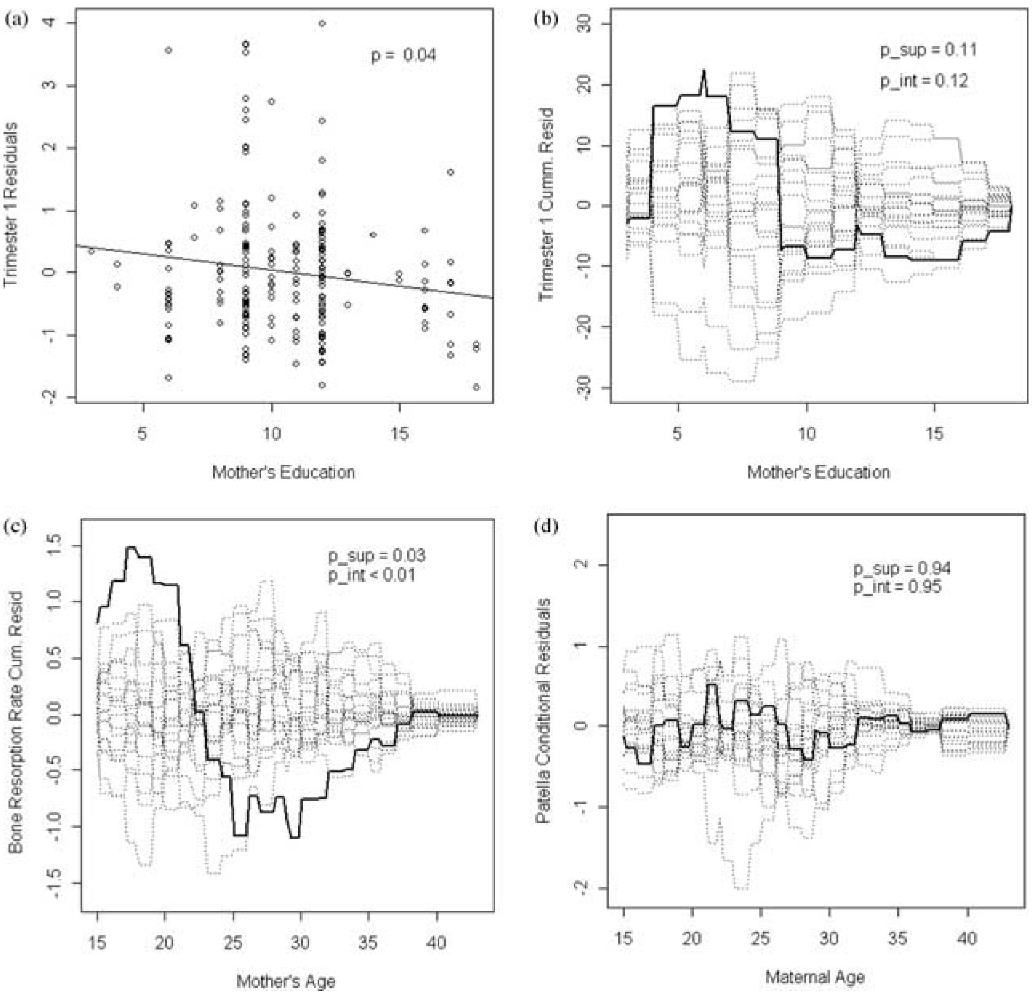

Figure 2. Residual diagnostics showing: (a) latent circulating lead residuals versus mother's education and regression-based p-value; (b) latent circulating lead residuals cumulatively summed with respect to mother's education and simulation-based p-values; (c) significant association between the bone resorption rate residuals and mother's age; (d) lack of association between the conditional residuals of patella lead concentration and age. Cumulative sum plots include simulated cumulative sums under the null hypothesis (light lines).

variable was assessed by using a nonlinear link function $g(\cdot)$ in the structural part of the model

$$\textbf{Model 2:}\ \ g(U_{1i}) = \gamma_{11}Z_{1i} + \gamma_{12}Z_{2i} + \zeta_{1i}.$$

To evaluate the power of the test of the functional form of a covariate on a surrogate, we considered two alternatives for the measurement part of the model

$$\textbf{Model 3a:}\ \ X_{3i} = \nu_3 + \lambda_3 U_{1i} + \kappa Z_{1i} + \epsilon_{3i},$$

$$\textbf{Model 3b:}\ \ X_{3i} = \nu_3 + \lambda_3 U_{1i} + \lambda U_{1i}^2 + \epsilon_{3i}.$$

Given that the assumed model does not include covariates in the measurement part of the model, the misspecification represented by Model 3a corresponds to item bias. Similar to Model 1c, data on Z_1 is used to construct the diagnostic. Model 3b corresponds to a nonlinear association between the latent variable and the observed surrogate. In this case, U_1 is already in the model, but the diagnostic still uses an estimate of $U_1, E(U_1)$, to construct the cumulative sum. Finally, to investigate the power of the test of link between a surrogate and its predictors, we considered the following alternative for the measurement model

$$\textbf{Model 4:}\ \ g(X_{3i}) = \nu_3 + \lambda_3 U_{1i} + \epsilon_{3i},$$

where $g(\cdot)$ is a nonlinear link function.

Tables 3 and 4 show the empirical sizes and power for the linearity tests considered. The empirical sizes of the tests based on the cumulative sums of residuals are within Monte Carlo error of the nominal size. The cumulative sum tests based on the latent variable residuals enjoy reasonable power in larger sample sizes, although their power is reduced with increasing measurement error variance in the surrogates. In contrast to linear mixed models, the link tests had little power to detect misspecified functional forms of covariates (Pan and Lin, 2005); the power never exceeded 30% (not shown). The test evaluating the functional form of the covariates on either the latent or surrogate variables have good power. Under all scenarios, the CVM-type test $\Im_2(\cdot)$ is more powerful than the KS-type test $\Im_\infty(\cdot)$. The normality test, as expected,

Table 3

Empirical size and power for cumulative latent variable residuals tests, nominal size 0.05

		Surrogates with 30% meas. error				Surrogates with 50% meas. error			
	n	$\Im_\infty(\cdot)$	$\Im_2(\cdot)$	(Wald)	CVM Test ζ_1	$\Im_\infty(\cdot)$	$\Im_2(\cdot)$	(Wald)	CVM Test ζ_1
Correct Model									
Covariate Test	200	0.05	0.05			0.042	0.043		
	300	0.06	0.05			0.043	0.046		
Link Test	200	0.06	0.06			0.05	0.048		
	300	0.05	0.05			0.047	0.044		
Model 1a									
$\gamma_{13} = 0.25$	200	0.33	0.54	(0.95)	0.06	0.28	0.42	(0.93)	0.04
	300	0.60	0.90	(>0.99)	0.05	0.48	0.79	(>0.99)	0.06
$\gamma_{13} = 0.50$	200	0.92	0.99	(>0.99)	0.11	0.79	0.97	(>0.99)	0.09
	300	0.99	0.99	(>0.99)	0.24	0.99	>0.99	(>0.99)	0.15
Model 1b									
$\gamma_{12} = 0.25$	200	0.14	0.20		0.056	0.12	0.15		0.06
	300	0.16	0.30		0.051	0.24	0.40		0.06
$\gamma_{12} = 0.50$	200	0.46	0.72		0.061	0.31	0.49		0.06
	300	0.90	0.99		0.087	0.82	0.99		0.08
Model 1c									
$\gamma_{13} = 0.5$	200	0.34	0.72	(>0.99)	0.051	0.52	0.92	(>0.99)	0.05
	300	0.85	>0.99	(>0.99)	0.05	>0.99	0.99	(>0.99)	0.05
$\gamma_{13} = 1.0$	200	0.75	>0.99	(>0.99)	0.05	0.99	>0.99	(>0.99)	0.05
	300	0.99	>0.99	(>0.99)	0.051	>0.99	>0.99	(>0.99)	0.06
Model 2									
$g(\cdot) = \log(\cdot)$	200	0.14	0.46		0.85	0.18	0.43		0.99
	300	0.40	0.69		0.95	0.41	0.65		0.94
$g(\cdot) = \sqrt{\cdot}$	200	0.48	0.57		0.99	0.45	0.54		>0.99
	300	0.67	0.71		0.92	0.64	0.68		>0.99

detected misspecified means in some settings. The power of the Wald test is greater, but the test requires a correctly specified, nested alternative.

5. Example: Fetal Lead Exposure

We use SEMs to study the relationships between bone- and blood-based biomarkers of fetal lead exposure and apply the diagnostic tools defined in Section 3 to this model. The model, represented in Figure 1, succinctly describes the associations of interest and utilizes all available data, as opposed to conducting multiple analyses with traditional regression methods. Table 1 summarizes the abbreviations used in the following algebraic form of the model.

5.1 *Measurement Model*

We relate the blood-based biomarkers to latent circulating lead as follows. At time $t, t = 0, 1, 2, 3$ (before pregnancy, and trimesters 1, 2, 3) let,

$X_{1t} = U_t + \epsilon_{1t}$	Model for Plasma Lead
$X_{2t} = \nu_{2t} + \lambda_2 U_t + \epsilon_{2t}$	Model for Blood Lead, ABC Laboratory
$X_{3t} = \nu_{3t} + \lambda_3 U_t + \epsilon_{3t}$	Model for Blood Lead, Smith Laboratory
$X_{43} = \nu_{43} + \lambda_{43} U_3 + \epsilon_{43}$	Model for Cord Blood Lead,

where U_t represents circulating lead at time t. Circulating lead takes the units of plasma lead, both in terms of its scale as well as central tendency because $\lambda_1 = 1$ and $\nu_{1t} = 0$. The model is weakly time invariant (Bollen, 1989) because the only parameters not allowed to vary over time are the factor loadings, λ—i.e., the associations between the scales of the latent and observed variables do not change over time. To account for laboratory effects in the blood lead concentrations, we allow $\boldsymbol{\epsilon}_2 = (\epsilon_{20}, \epsilon_{21}, \epsilon_{22}, \epsilon_{23})^T$ and $\boldsymbol{\epsilon}_3 = (\epsilon_{30}, \epsilon_{31}, \epsilon_{32}, \epsilon_{33})^T$ to have banded correlation structures, but $\boldsymbol{\epsilon}_2$ and $\boldsymbol{\epsilon}_3$ are assumed to be independent; i.e., for $j = 2, 3, \text{corr}(\epsilon_{jt}, \epsilon_{j;t+k}) \neq 0$ for $k = 1$, but is 0 for $k > 1$, and $\text{corr}(\epsilon_{jt}; \epsilon_{j't'}) = 0$ for $j \neq j'$.

Bone resorption rates, measured at $t = 1, 2, 3$, are modeled using a mixed effects model, $X_{t4} = U_4 + \kappa t + \epsilon_{t4}$, where $U_4 \sim \text{Normal}(\alpha_4, \psi_4)$ represents a random intercept, and κ is a fixed time effect. The random intercept can be interpreted as the mother's intrinsic resorption rates. The model for bone lead concentrations is

$X_{15} = U_5 + \epsilon_{15}$	Patella lead, before pregnancy
$X_{25} = \nu_{25} + \lambda_{25} U_5 + \epsilon_{25}$	Tibia lead, before pregnancy
$X_{35} = \nu_{35} + U_5 + \epsilon_{35}$	Patella lead, 1 mo. post partum
$X_{45} = \nu_{45} + \lambda_{25} U_5 + \epsilon_{45}$	Tibia lead, 1 mo. post partum,

which assumes that bone lead burden, U_5, is measured in units of patella lead concentration. The population average for bone lead burden is assumed to be equal to the average observed for patella lead before pregnancy. After pregnancy this concentration is, on average, ν_{35} units different from that before pregnancy (i.e., fixed time effect). Tibia lead concentration is

Table 4
Empirical size and power for cumulative conditional residuals tests, nominal size 0.05

		Surrogates with 30% meas. error				Surrogates with 50% meas. error			
	n	$\Im_\infty(\cdot)$	$\Im_2(\cdot)$	(Wald)	CVM Test ϵ_3	$\Im_\infty(\cdot)$	$\Im_2(\cdot)$	(Wald)	CVM Test ϵ_3
Correct Model									
Covariate Test	200	0.059	0.059			0.040	0.048		
	300	0.050	0.050			0.044	0.042		
Covariate Test[a]	200	0.041	0.041			0.039	0.044		
	300	0.056	0.059			0.041	0.042		
Link Test	200	0.050	0.041			0.042	0.046		
	300	0.059	0.050			0.041	0.043		
Model 3a									
$\kappa = 0.5$	200	0.13	0.38	(0.34)	0.054	0.08	0.12	(0.19)	0.07
	300	0.25	0.66	(0.51)	0.052	0.08	0.18	(0.32)	0.06
$\kappa = 1.0$	200	0.42	0.93	(0.90)	0.057	0.12	0.28	(0.72)	0.05
	300	0.74	0.99	(0.99)	0.052	0.19	0.52	(0.94)	0.05
Model 3b[a]									
$\lambda = 0.25$	200	0.31	0.53	(>0.99)	0.18	0.16	0.25	(0.97)	0.07
	300	0.52	0.78	(>0.99)	0.31	0.24	0.40	(>0.99)	0.11
$\lambda = 0.50$	200	0.94	0.99	(>0.99)	0.94	0.46	0.77	(0.99)	0.53
	300	0.95	0.99	(>0.99)	0.99	0.70	0.95	(>0.99)	0.75
Model 4									
$g(\cdot) = \log(\cdot)$	200	0.11	0.23		0.99	0.08	0.08		0.95
	300	0.22	0.36		0.99	0.14	0.14		0.75
$g(\cdot) = \sqrt{\cdot}$	200	0.14	0.48		0.93	0.36	0.48		0.80
	300	0.83	0.96		>0.99	0.58	0.75		>0.99

[a] Uses $1\{E(U_{1i}) \leq t\}$ in the definition of the cumulative sum.

also a surrogate of bone lead burden; 1 unit of patella lead is roughly equivalent to λ_{25} units of tibia lead before and after pregnancy.

5.2 *Associations Between Latent Variables*

We impose a longitudinal model on latent circulating-lead exposure. That is, $U_t = \alpha_t + \beta_1 U_4 + \beta_2 U_5 + \gamma_1 Z_1 + \gamma_2 Z_2 + \gamma_3 Z_3 + \zeta_t$, where α_t is a trimester-specific mean (to avoid assumptions on the trend of the latent variables). The effects of bone resorption rates, U_4, and bone lead burden, U_5, on circulating lead are given by β_1 and β_2, respectively. The covariance of ζ_t, $t = 0, 1, 2, 3$ is modeled as a banded-heterogeneous matrix. The model defined for the blood-based lead biomarkers can be seen as a latent variable model for longitudinal data with multiple continuous outcomes (Roy and Lin, 2000), except that here latent variables are also predictors. Finally, we model the effects of covariates on bone lead burden: $U_5 = \alpha_5 + \gamma_{15} Z_1 + \gamma_{25} Z_2 + \gamma_{35} Z_3 + \zeta_5$, where ζ_5 is assumed to be independent from ζ_4 and from ζ_t, $t = 0, 1, 2, 3$.

5.3 *Diagnostics*

Given the variety of residuals and available tests, a strategy for the order in which model assumptions are evaluated is needed. Since the distribution tests require correct mean and variance specification, we apply linearity tests first. Specifically, we use linearity diagnostics to detect whether covariates are missing from the model misspecification, or whether linear relationships are incorrectly specified. We then inspect distributional assumptions by first using the omnibus test for normality, then checking the distributions of the latent variable residuals, and, finally, examining the conditional residuals. Here we show a sample of the diagnostic results to illustrate the use of the methods.

After fitting a model, it may be of interest to assess whether additional covariates should be added to it. For instance, we considered adding maternal education as a predictor of latent circulating lead exposure. A sensible approach may be to plot the circulating lead exposure residuals against education and fit a regression line (Figure 2a). In this case, the p-value obtained from a regression of the residuals against maternal education suggests that education should be added as a predictor. However, inferences from such an approach may be problematic as they do not adjust for the estimation of $\boldsymbol{\theta}$. Thus, we analyze the cumulative sum of the circulating lead residuals with respect to education (Figure 2b). Although the observed cumulative sum appears to deviate from the simulated sums (light lines), objective inference based on the simulated p-value does not call for adding education as a predictor. Nevertheless, we fitted a model including education as a predictor of trimester 1 exposure. From the corresponding Wald test, we concluded education was not a statistically significant predictor of circulating lead exposure.

Figure 2c displays the cumulative sum of latent resorption rate residuals with respect to maternal age. The observed cumulative sum clearly stands out from the simulated sums. The drastic increase in the sum indicates a preponderance of positive residuals at younger ages, and the sharp decrease indicates an excess of negative residuals at older ages. This pattern in the residuals and their cumulative sum is consistent with a missing linear term for age (Z_1), and thus suggests modeling: $U_4 = \alpha_4 + \gamma_{14} Z_1 + \zeta_4$. We included the term $\gamma_{14} Z_1$

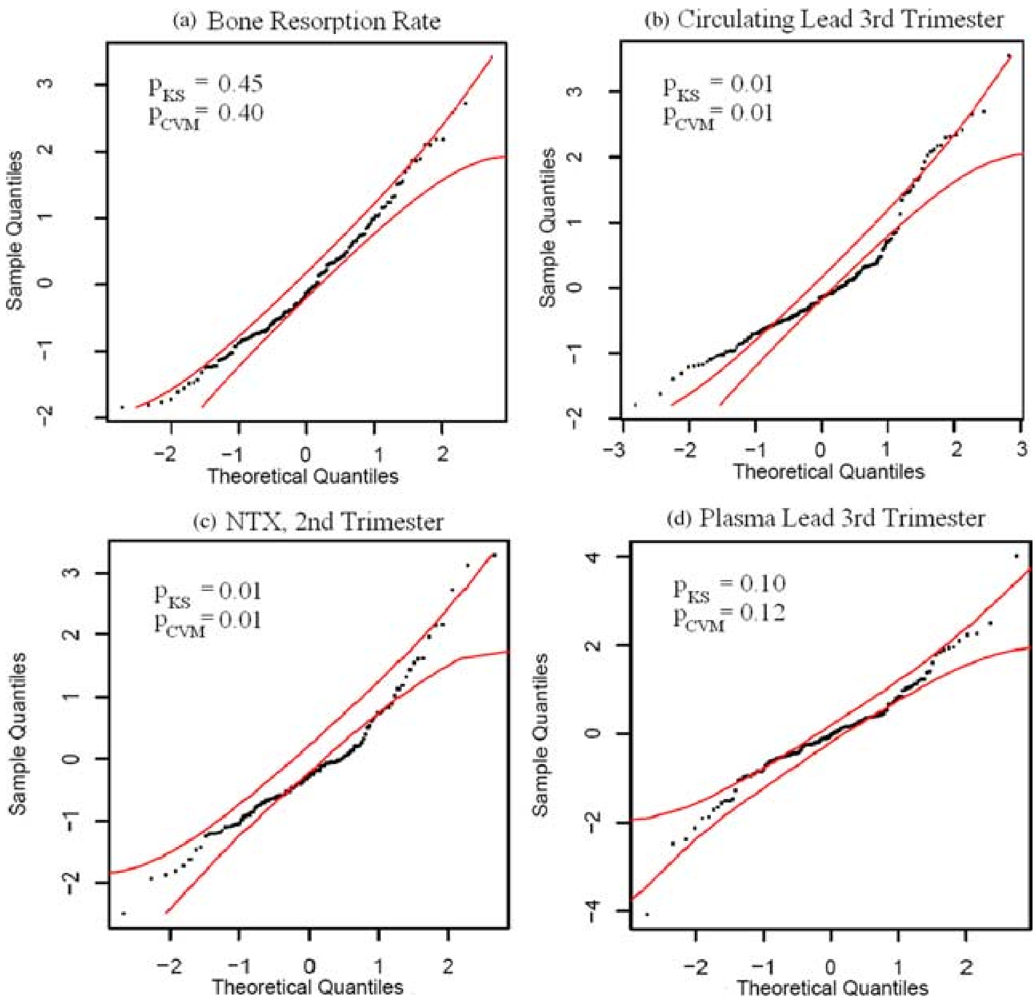

Figure 3. *Q*–*Q* plots with point by point 95% confidence intervals and simulation based p-values corrected for parameter estimation showing (a) no significant lack of fit for bone resorption rate residuals; (b) lack of normality for the circulating lead residuals; (c) significant lack of normality for bone resorption biomarker (NTX) at trimester 2; (d) marginal lack of normality for plasma lead residuals at trimester 3.

in the model, and again calculated the cumulative sum of the bone resorption residuals with respect to age. From the updated plot and test, we concluded there was no significant association between the bone resorption rate residuals and maternal age, thereby suggesting that a linear term may be sufficient in modeling the bone resorption rate-age association.

Figure 2d presents the cumulative sum of conditional residuals for patella lead concentration. The observed sum is well within the cloud of simulated sums. Thus, no evidence exists of a misspecified association between patella lead concentration and age. Similarly, cumulatively summing the residuals of plasma lead concentrations gave no evidence of a misspecified links between the plasma biomarkers and its predictors (not shown).

Next we examine distributional assumptions. We first conducted an omnibus test of normality based on the standardized marginal residuals, and concluded significant lack of normality. However, the omnibus test provides no indication of which residuals $\boldsymbol{\zeta}$ or $\boldsymbol{\epsilon}$ violate the normality assumption. Therefore, we examined the conditional and latent components of the error separately. Figures 3a and b show *Q*–*Q* plots for two of the six latent variable residuals in the model, point-wise confidence intervals calculated using the variance estimator proposed by Houseman et al. (2004), and p-values for the KS and CVM type tests. There was no violation of normality in the distribution latent variable residuals for bone resorption rates. In contrast, the distributions of (latent) circulating lead residuals violate normality assumptions. The shape of the *Q*–*Q* plot would instead suggest a left-skewed distribution.

Next, we inspected the conditional error distributions. The conditional residuals for both bone lead concentrations (before and after) and blood lead concentrations (both laboratories, at every trimester) did not deviate significantly from normality. The bone resorption concentrations reveal deviation from normality at trimester 2 (Figure 3c). The conditional residuals for plasma lead show marginally significant lack of normality (Figure 3d); the Q–Q plot is indicative of a heavy tailed error distribution. Thus, a t or logistic distribution might instead be considered for these conditional errors. Note that these figures show the specificity of the normality tests; that is, the latent variable residuals for the bone resorption rates (Figure 3a) do not appear to violate normality assumptions, whereas one of its surrogates does (NTX, T2, Figure 3c).

6. Conclusions

We proposed and implemented residual-based diagnostics that assess distributional and linearity assumptions commonly made in classical SEMs. The diagnostics are based on three types of residuals that isolate the various sources of error in the model, namely, marginal, conditional, and latent variable residuals. The theoretical basis for making inferences based on the proposed diagnostics follows from research on residual diagnostics for linear mixed models (Houseman et al., 2004, 2006; Pan and Lin, 2005). Both tests are adjusted for the estimation of all model parameters, although more simple adjustments for parameter estimation may be sufficient for the normality tests (see Web Appendix D). The proposed diagnostics improve upon available model specification tools for SEMs because they not only lend themselves to graphical displays, but are also based on individual-level residuals. Classical model specification tools, such as modification and χ^2 fit indices, are based on aggregate forms of the data (Beck, 1982), and require specific alternatives to test against.

Further research is needed on the sequence in which these diagnostics should be implemented; that is, there is possible confounding between various types of model misspecification (Eberly and Thackeray, 2005). We examined linearity assumptions before normality assumptions because distribution diagnostics assume correctly specified mean and variances; but diagnostics for linearity may not necessitate correctly specified error distributions (Pan and Lin, 2005; also see Web Appendix E).

Furthermore, because in SEMs incorrect covariance assumptions on the error terms leads to biased conditional mean parameter estimates (Sammel and Ryan, 2002) it is unclear whether testing the default linear associations in SEMs may also require correctly specified covariance structures for the errors. This is in contrast to diagnostics for linear mixed models, where correctly specified covariance matrices are not required to make inferences on the predictive part of the model (Pan and Lin, 2005). Thus, developing projection matrices based on robust variance matrices might be a promising next step. Developing methods to assess the covariance matrices of the errors might also be of interest.

7. Supplementary Materials

Web Appendices referenced in Sections 3 and 6, as well as the `R` code, are available under the Paper Information link at the Biometrics website `http://www.biometrics.tibs.org`

Acknowledgements

The authors would like to thank Prof Howard Hu for valuable discussions and the use of the data for this article. This publication was partially supported by a Predoctoral Fellowship from the Howard Hughes Medical Institute (HHMI) to the first author. It was supported also in part by a grant from the American Chemistry Council (ACC # 2843). It has not been formally reviewed by the HHMI or the ACC, and the views expressed in this document are solely those of the authors.

References

Arminger, G. and Schoenberg, R. J. (1989). Pseudo maximum-likelihood estimation and a test for misspecification in mean and covariance structure models. *Psychometrika* **54,** 409–425.

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., and Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association* **92,** 1375–1386.

Bauer, D. (2005). The role of non-linear factor-to-indicator relationships in tests of measurement invariance. *Psychological Methods* **10,** 305–316.

Beck, R. (ed.) (1982). *Handbook of Methods for Metecting Test Bias.* Baltimore: Johns Hopkins University Press.

Bollen, K. A. (1989). *Structural Equation Models with Latent Variables.* New York: Wiley.

Bollen, K. A. and Arminger, G. (1991). Observational residuals for factor analysis and structural equation models. *Sociological Methodology* **21,** 235–262.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance-structures. *British Journal of Mathematical and Statistical Psychology* **37,** 62–83.

Budtz-Joergensen, E., Keiding, N., Grandjean, P., Weihe, P., and White, R. F. (2002). Estimation of health effects of prenatal mercury exposure using structural equation models. *Environmental Health: A Global Access Science Source* **1,** 2.

Carroll, R. J., Ruppert, C., Crainiceanu, C. M., Tosteson, T. D., and Karagas, M. R. (2004). Nonlinear and nonparametric regression and instrumental variables. *Journal of the American Statistical Association* **99,** 736–750.

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society, Series B* **62,** 355–366.

Eberly, L. E. and Thackeray, L. M. (2005). On Lange and Ryan's plotting technique for diagnosing non-normality of random effects. *Statistics and Probability Letters* **75**(2), 77–85.

Houseman, E. A., Ryan, L. M., and Coull, B. A. (2004). Cholesky residuals for assessing normal errors in a linear model with correlated outcomes. *Journal of the American Statistical Association* **99,** 383–394.

Houseman, E. A., Coull, B. A., and Ryan, L. M. (2006). A functional-based distribution diagnostic for a linear model with correlated outcomes. *Biometrika* **93,** 911–926.

Lee, S. Y. and Lu, B. (2003). Case deletion diagnostics for nonlinear structural equation models. *Multivariate Behavioral Research* **38,** 375–400.

Lee, S. Y. and Song, X. Y. (2004). Maximum likelihood analysis of a general latent variable model with hierarchically mixed data. *Biometrics* **60,** 624–636.

Lee, S. Y. and Tang, N. S. (2004). Local influence analysis of nonlinear structural equation models. *Psychometrika* **69,** 573–592.

Lin, D. Y., Wei, L. J., and Ying, Z. (2002). Model-checking techniques based on cumulative residuals. *Biometrics* **58,** 1–12.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. New York: John Wiley.

Liu, X., Wall, M. M., and Hodges, J. S. (2005). Generalized spatial structural equation models. *Biostatistics* **6,** 539–557.

Longford, N. T. (1993). *Random Coefficient Models.* New York: Oxford University Press.

Muthén, L. K. and Muthén, B. O. (1998–2004). *Mplus User's Guide,* 3rd edition. Los Angeles, CA: Muthen & Muthen.

Pan, Z. and Lin, D. Y. (2005). Goodness-of-fit methods for generalized linear mixed models. *Biometrics* **61,** 1000–1009.

Proust, C., Jacqmin-Gadda, H., Taylor, J. M. G., Ganiayre, J., and Commenges, D. (2006). A nonlinear model with latent process for cognitive evolution using multivariate longitudinal data. *Biometrics* **62,** 1014–1024.

Roy, J. and Lin, X. (2000). Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics* **56,** 1047–1054.

Sammel, M. D. and Ryan, L. M. (2002). Effects of covariance misspecification in a latent variable model for multiple outcomes. *Statistica Sinica* **12,** 1207–1222.

Sánchez, B. N., Budtz-Joergensen, E., Ryan, L. M., and Hu, H. (2005). Structural equation models: A review with applications to environmental epidemiology. *Journal of the American Statistical Association* **100,** 1443–1455.

Su, J. Q. and Wei, L. J. (1991). A lack-of-fit test for the mean function in a generalized linear model. *Journal of the American Statistical Association* **86,** 420–426.

Tellez-Rojo, M. M., Hernandez-Avila, M., Lamadrid-Figueroa, H., Smith, D., Hernandez-Cadena, L., Mercado, A., Aro, A., Schwartz, J., and Hu, H. (2004). Impact of bone lead and bone resorption on plasma and whole blood lead levels during pregnancy. *American Journal of Epidemiology* **160,** 668–678.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge, U.K.: Cambridge University Press.

Wall, M. M. and Amemiya, Y. (2000). Estimation for polynomial structural equation models. *Journal of the American Statistical Association* **95,** 929–940.

Received February 2007. *Revised January* 2008.
Accepted January 2008.