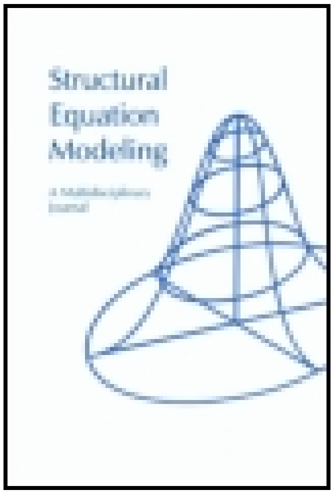

Structural Equation Modeling: A Multidisciplinary Journal
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hsem20

On the Performance of Likelihood-Based Difference Tests in Nonlinear Structural Equation Models

Carla Gerhard[a], Andreas G. Klein[a], Karin Schermelleh-Engel[a], Helfried Moosbrugger[a], Jana Gäde[a] & Holger Brandt[a]

[a] Goethe University, Frankfurt, Germany

Published online: 08 Oct 2014.

Structural Equation Modeling: A Multidisciplinary Journal, 00: 1–12, 2014

ISSN: 1070-5511 print / 1532-8007 online
DOI: 10.1080/10705511.2014.935752

Routledge
Taylor & Francis Group

On the Performance of Likelihood-Based Difference Tests in Nonlinear Structural Equation Models

Carla Gerhard, Andreas G. Klein, Karin Schermelleh-Engel, Helfried Moosbrugger, Jana Gäde, and Holger Brandt
Goethe University, Frankfurt, Germany

This article investigates likelihood-based difference statistics for testing nonlinear effects in structural equation modeling using the latent moderated structural equations (LMS) approach. In addition to the standard difference statistic T_{D}, 2 robust statistics have been developed in the literature to ensure valid results under the conditions of nonnormality or small sample sizes: the robust T_{DR} and the "strictly positive" T_{DRP}. These robust statistics have not been examined in combination with LMS yet. In 2 Monte Carlo studies we investigate the performance of these methods for testing quadratic or interaction effects subject to different sources of nonnormality, nonnormality due to the nonlinear terms, and nonnormality due to the distribution of the predictor variables. The results indicate that T_{D} is preferable to both T_{DR} and T_{DRP}. Under the condition of strong nonlinear effects and nonnormal predictors, T_{DR} often produced negative differences and T_{DRP} showed no desirable power.

Keywords: interaction effects, likelihood-ratio test, Monte Carlo study, nonlinear SEM, robust difference test, strictly positive difference test

Structural equation modeling (SEM) is a common statistical tool for modeling relationships between latent variables that cannot be measured without errors. In SEM, latent exogenous and endogenous variables are operationalized by observable indicators, which allow for measurement errors. Besides the use of linear SEM, nonlinear SEM has gained more and more attention over the last decade and it has frequently been used in the context of applied behavioral and social science research (cf. Kline, 2010; Marsh, Wen, Nagengast, & Hau, 2012; Moosbrugger, Schermelleh-Engel, Kelava, & Klein, 2009). Nonlinear SEM enhances ordinary SEM by adding latent interaction, quadratic terms, or both to the structural equation.

For the estimation of nonlinear SEM, Klein and Moosbrugger (2000) developed a maximum likelihood estimation method for the first time, the latent moderated structural equations (LMS) method. LMS is a distribution-analytic method that takes the nonnormality of the criterion variable (η)[1] into account by analyzing its distribution as a mixture of several conditionally normal distributions (cf. Klein & Moosbrugger, 2000; see also Kelava et al., 2011). By this, only the predictor variables, the measurement errors, and the residual variable (x, ξ, δ, ε, ζ) are required to follow a multivariate normal distribution. LMS includes a model difference test T_{D} for latent interaction effects. Previous simulation studies suggest that T_{D} might perform well under ideal simulation conditions (Cham, West, Ma, & Aiken, 2012; Klein, 2000; Klein & Moosbrugger, 2000; Klein & Muthén, 2007).

Nonlinear terms in nonlinear SEM are generally nonnormally distributed (cf. Klein & Moosbrugger, 2000; Moosbrugger, Schermelleh-Engel, & Klein, 1997). Consequently, the latent endogenous variables are typically not normally distributed either. Also, there might be additional nonnormality due to nonnormal predictor variables. Because of these two different sources of nonnormality, the assumption of multivariate normally distributed indicators

Correspondence should be addressed to Carla Gerhard, Department of Psychology, Goethe-Universität Frankfurt, Campus Westend—PEG, Grüneburgplatz 1, D-60323, Frankfurt am Main, Germany. E-mail: Gerhard@psych.uni-frankfurt.de

[1]In the following, we will use the LISREL notation (cf. Jöreskog & Yang, 1996).

could often be violated in nonlinear SEM. This raises the question as to what degree the standard model difference test T_{D} still produces reliable results when distributional assumptions about the predictors are violated.

In the literature, two robust test statistics designed for nonnormal data have been proposed by Satorra and Bentler (2001, 2010): the Satorra–Bentler (SB) scaled difference chi-square statistic T_{DR} and the strictly positive SB scaled difference chi-square statistic T_{DRP}. Up to the present neither of these two statistics has been examined for use with LMS. It is yet unknown if these robust methods have any benefit for detecting nonlinear effects. The goal of this article is to present the theoretical background, the features, and the application of standard and robust likelihood-ratio tests for testing single or several nonlinear effects in latent nonlinear SEM. We systematically investigate the standard LMS difference test T_{D} and the robust difference tests T_{DR} and T_{DRP}.

In empirical research it is often of special interest to test not only interaction effects, but also quadratic effects or several nonlinear effects simultaneously. Because until now T_{D} has only been investigated for testing interaction effects, this article focuses on the detection of quadratic effects and on the detection of quadratic and interaction effects simultaneously. For this purpose, different types of nested models were compared in two Monte Carlo studies. In addition, different sources of nonnormality were simulated. The first Monte Carlo study investigated the influence of nonnormality due to nonlinear terms and the second Monte Carlo study investigated the influence of additional nonnormality due to nonnormally distributed predictors.

In the following, different ways of performing chi-square difference tests in linear SEM and the specifics of likelihood-based difference tests for LMS in nonlinear model structures are outlined.

REVIEW OF STANDARD CHI-SQUARE AND LIKELIHOOD-BASED DIFFERENCE TEST

To test the significance of parameters of nested models, one can test a target model against a more restricted baseline model. In *linear SEM,* the chi-square difference statistic T_{D} compares the chi-square values of two nested SEM models. Nested models are equivalent except for a subset of free parameters in one model that is fixed in the second model. In the more restrictive nested model (M_0), one or several parameters are fixed to zero. In the less restrictive model (M_1) these additional parameters are estimated. The difference statistic T_{D} calculates the difference between the chi-squares χ_0^2 and χ_1^2 for M_0 and M_1, and it is chi-square distributed with $df_D = df_0 - df_1$ degrees of freedom (Steiger, Shapiro, & Browne, 1985):

$$T_D = \chi_D^2 = \chi_0^2 - \chi_1^2. \tag{1}$$

For *nonlinear SEM,* it is not possible so far to obtain reliable chi-square statistics. However, it is possible to test single or multiple nonlinear effects by a standard likelihood-ratio test. The difference between the log-likelihood LL_0 of a more restrictive model (M_0) and the log-likelihood LL_1 of a less restrictive model (M_1) multiplied by –2 is chi-square distributed (Bollen, 1989; cf. Muthén & Muthén, 2006):

$$T_D = -2(LL_0 - LL_1). \tag{2}$$

The number of degrees of freedom for the difference statistic T_{D} is calculated by subtracting the number of estimated parameters of the more restrictive model (p_0) from the number of estimated parameters of the less restrictive model (p_1): $df_{\mathrm{D}} = p_1 - p_0$. Thus, the likelihood-based difference test T_{D} compares the log-likelihoods of the nested models, with greater differences reducing the probability that the more restrictive M_0 model is retained.

Previous studies report a good power of T_{D} for detecting interaction effects (Klein, 2000; Klein & Moosbrugger, 2000; Klein & Muthén, 2007). These findings suggest that T_{D} performs well under ideal simulation conditions. These studies focused on the detection of interaction effects and did not consider the detection of quadratic effects or different types of nonlinear effects simultaneously. Nevertheless, it was concluded that standard difference tests might perform well when the interaction effect is sufficiently large.

Additional support for using the standard likelihood-ratio test in nonlinear SEM is given by a study of Klein and Moosbrugger (2000), where the authors took a closer look at the distribution of T_{D}. In a Monte Carlo study they evaluated the distribution of T_{D} for a latent moderated model with normally distributed predictor variables and sample size $N = 400$ and compared it to the theoretical chi-square distribution. They reported that the distribution of the test statistic T_{D} did not deviate significantly from the theoretical distribution.

Robust Chi-Square and Likelihood-Based Difference Test

The calculation of the robust difference statistic T_{DR} and the strictly positive test statistic T_{DRP} given by Satorra and Bentler (2001, 2010) is based on the chi-square values of model fit for linear SEM. The regular chi-square of model fit statistic T of a linear structural equation model is calculated under the assumption that the variables are normally distributed. If this is not the case, Satorra and Bentler showed that T is asymptotically distributed as a mixture of chi-square distributions of 1 *df.* Therefore a robustified version T_{R} of the chi-square values was proposed to improve the chi-square approximation in nonasymptotic and nonnormal applications (Satorra & Bentler, 1994). T_{R} approximates the expected chi-square distribution under conditions of nonnormality by estimating a correction factor $\hat{c}$ to adjust the mean and variance of the chi-square values (Satorra & Bentler, 1994). For general types of distribution, a corrected statistic T_{R} was proposed, wherein T is divided by $\hat{c}$. The correction factor $\hat{c}$ is calculated from the estimated asymptotic covariance matrix of the sample covariances and variances, which contains information about higher order moments, such as skewness and kurtosis. The value of $\hat{c}$ is a function of the observed multivariate kurtosis and the degrees of freedom of the model. The chi-square values, which are biased due to the given nonnormality, are scaled so that the mean of T_{R} matches again with the degrees of freedom, $E(T_R) = df$ (Satorra, 1990, 1992; Satorra & Bentler, 1994; see also Curran, West, & Finch, 1996). The robust SB statistic T_{R} has been well studied under various conditions of nonnormality and it appears to perform well under nonnormal conditions, even in small samples (e.g., Chou, Bentler, & Satorra, 1991; Curran et al., 1996; Hu, Bentler, & Kano, 1992).

The difference between two T_{R} statistics of two nested models is not chi-square distributed anymore (Satorra, 2000). For such situations Satorra and Bentler (2001) developed a method to calculate robust chi-square difference tests by estimating an additional correction factor $\hat{c}_D$ that is based on the correction factor $\hat{c}_0$ of the M_0 model and on the correction factor $\hat{c}_1$ of the M_1 model, as well as on the degrees of freedom of the nested models (cf. Satorra & Bentler, 2001):

$$\hat{c}_D = (df_0\hat{c}_0 - df_1\hat{c}_1)/(df_0 - df_1). \tag{3}$$

To obtain the robust chi-square difference statistic T_{DR}, T_{D} is divided by $\hat{c}_D$:

$$T_{DR} = T_D/\hat{c}_D = (\chi_0^2 - \chi_1^2)/\hat{c}_D. \tag{4}$$

The performance of T_{DR} is less well investigated than the performance of T_{R}. Under nonnormal conditions for linear SEM models, the robust statistics are recommended for both the chi-square test of model fit and the chi-square difference test (Satorra & Bentler, 1994, 2001).

For nonlinear SEM the robust difference test has only been investigated for product indicator (PI) approaches (Cham et al., 2012), for which the chi-square values are provided in the software output. The robust difference tests for PI approaches are not directly comparable to those for LMS difference tests. For LMS, chi-square values of fit are not available. The reason for this lies in the fact that for nonlinear SEM a statistic for a saturated model cannot be computed in a straightforward manner (Klein & Schermelleh-Engel, 2010). A common misconception about the robust difference statistic developed by Satorra and Bentler is that this statistic could only be applied when the chi-square values of model fit are available. This is not the case because likelihood values are sufficient for its computation. When neither the chi-square values of the model fit nor the number of degrees of freedom for the models are available, it is suggested to use the robust difference test statistic T_{DR} by using the log-likelihood values instead of the chi-square values (Asparouhov & Muthén, 2012; Muthén & Muthén, 2006). For a calculation of the robust likelihood-based difference test one needs the log-likelihood values (LL_0, LL_1), the number of estimated parameters (p_0, p_1), and the estimated correction factors ($\hat{c}_0$, $\hat{c}_1$) for the nested models. Instead of using Equation 3 an alternative formula for $\hat{c}_D$ is applied (Asparouhov & Muthén, 2012; Muthén & Muthén, 2006), which is given by

$$\hat{c}_D = \left(p_0\hat{c}_0 - p_1\hat{c}_1\right) / \left(p_0 - p_1\right). \tag{5}$$

Considering Equations 2 and 5, the likelihood-based T_{DR} (Asparouhov & Muthén, 2012; Muthén & Muthén, 2006) is then given by

$$T_{DR} = T_D/\hat{c}_D = -2(LL_0 - LL_1)/\hat{c}_D. \tag{6}$$

In nonlinear SEM, M*plus* (Muthén & Muthén, 1998–2010) provides the log-likelihood values and the correction factors for the nested models when the MLR (robust maximum likelihood) estimation option is used. It is the only estimation option for nonlinear SEM in M*plus* for which the required correction factors are available when using LMS. All necessary information for robust difference test statistic T_{DR} is provided, and it is explicitly recommended to perform the robust difference test when using MLR (Muthén & Muthén, 2010). As the required information is provided and T_{DR} is recommended, T_{DR} has often been applied in practice when using MLR (cf. Dimitruk, Schermelleh-Engel, Kelava, & Moosbrugger, 2007; Hughes & Kwok, 2006; Marshall, Miles, & Stewart, 2010; Willoughby, Cadigan, Burchinal, & Skinner, 2008).

When the MLR estimator for nonlinear SEM is used in the M*plus* program, it is tailored in particular for nonlinear SEM. A maximum likelihood estimator with a numerical integration algorithm that is based on LMS (Klein & Moosbrugger, 2000) is applied. Theoretically the nonnormality due to the nonlinear terms is already considered when using LMS for nonlinear SEM and an additional correction of the difference statistic might not make much of a difference as long as the predictor variables and the measurement errors themselves are normally distributed. However, the use of the correction formula for the LMS-based test might be of practical value when LMS is applied under less than optimal conditions where the distributional assumptions are violated. The usefulness of such a robustified version of the LMS difference test has not been examined yet.

Strictly Positive Robust Chi-Square and Likelihood-Based Difference Tests

Under certain conditions, the robust chi-square difference test might produce a negative T_{DR} value in linear SEM

(Bryant & Satorra, 2012; Satorra & Bentler, 2010). Negative chi-square values are improper and cannot be used. A negative T_{DR} occurs when $\hat{c}_D$ is negative. This is the case when $df_0\hat{c}_0 - df_1\hat{c}_1 < 0$, as the denominator $df_0 - df_1$ in Equation 3 is always positive. To solve this problem, a new adjustment of the correction method has been developed, the strictly positive robust difference statistic T_{DRP} (Satorra & Bentler, 2010). T_{DRP} has been recommended whenever negative difference values appear for the T_{DR} (Asparouhov & Muthén, 2012).

The idea of T_{DRP} is to estimate a third model M_{10}, which has the same model structure as M_1, but with the parameter estimates that occurred in the restricted M_0 model held fixed. The auxiliary model M_{10} can be seen as a nonoptimized version of M_1 (Asparouhov & Muthén, 2012). For the strictly positive chi-square difference test, M_{10} is estimated to obtain a new correction factor $\hat{c}_{10}$. The correction factor $\hat{c}_{10}$ is then used instead of $\hat{c}_1$ (cf. Equation 3) for a calculation of the strictly positive correction factor $\hat{c}_{DP}$:[2]

$$\hat{c}_{DP} = (df_0\hat{c}_0 - df_1\hat{c}_{10})/(df_0 - df_1). \tag{7}$$

The formula for the strictly positive difference statistic T_{DRP} is then given by

$$T_{DRP} = T_D/\hat{c}_{DP}. \tag{8}$$

The structure of M_{10} is the same as for M_1, but the parameter estimates of M_{10} have not been optimized for convergence and are identical with the parameter estimates of M_0.

The strictly positive difference test could be applied for nonlinear SEM as well to avoid negative differences. If the chi-square values and the number of degrees of freedom are not available, as is the case for LMS, an alternative likelihood-based formula was proposed by Asparouhov and Muthén (2012). Similar to Equation 7, and instead of $\hat{c}_1$ in Equation 5, the correction factor $\hat{c}_{10}$ of the nonoptimized M_1 model is used to calculate the strictly positive correction factor $\hat{c}_{DP}$ by

$$\hat{c}_{DP} = \left(p_0\hat{c}_0 - p_1\hat{c}_{10}\right) / \left(p_0 - p_1\right). \tag{9}$$

The strictly positive approach ensures that $\hat{c}_{10}$ is greater than $\hat{c}_0$, and that the difference values are positive. The likelihood-based T_{DRP} is then given by

$$T_{DRP} = T_D/\hat{c}_{DP} = -2(LL_0 - LL_1)/\hat{c}_{DP}. \tag{10}$$

The strictly positive difference test has not been examined for nonlinear SEM yet.

[2]For further details concerning the calculation of strictly positive difference test statistics in different SEM software packages, see Bryant and Satorra (2012).

RESEARCH QUESTIONS

In the following we want to address whether T_{D} is a suitable method for detecting single nonlinear effects or different nonlinear effects modeled simultaneously. Furthermore, as the multivariate distribution of the variables in nonlinear SEM is always nonnormal, the question arises if a robust test statistic might lead to an improvement in detecting nonlinear effects compared to the standard T_{D}. From the theoretical point of view, one might suggest that a correction of T_{D} for difference tests in nonlinear SEM is not necessarily required. As the nonlinearity of the data is inherent in both likelihood values of the compared models, the difference of these likelihoods might compensate the bias due to the nonnormality caused by nonlinear terms.

Using the LMS approach, the nonnormality in the data due to the nonlinear terms is already considered and a robust statistic might not be necessary as long as the predictor variables themselves are normally distributed. In view of these considerations, another important issue relates to possible additional nonnormality due to the predictor variables. There is some evidence that a severe violation of the normality assumption for the predictor variables might inflate the Type I error of standard LMS difference testing in nonlinear SEM (Cham et al., 2012; Klein, 2000; Klein & Moosbrugger, 2000). As in psychological and social science research, nonnormal predictors and small samples are not uncommon (cf. Micceri, 1989), so the question arises, what improvements might be achieved when performing robust likelihood-based difference tests under these nonnormal conditions. The robust T_{DR} might result in negative differences. There are no extensive simulation studies on this problem, but it has been noted in the context of linear SEM that this problem occurs especially in small samples or when the more restrictive model is highly incorrect (Satorra & Bentler, 2010). As negative difference values have already been reported in linear SEM, the strictly positive robust difference T_{DRP} is examined in this article, too.

Until now, neither the performance of T_{DR} nor the performance of T_{DRP} has been investigated in combination with LMS. It is unknown under what conditions the T_{DR} or the T_{DRP} might have an advantage over the standard test statistic T_{D}. To investigate the robustness of T_{D}, T_{DR}, and T_{DRP} under conditions relevant for empirical research, an extensive and comparative simulation study appears necessary. Therefore, for this article two Monte Carlo studies were conducted. The first study focused on nonnormality due to the nonlinear terms, where the size of the nonlinear effects was varied. The second study focused on additional nonnormality due to nonnormally distributed predictors, where the degree of nonnormality of the predictor variables was varied. The following questions are addressed:

1. What type of likelihood-based difference test (T_D, T_{DR}, and T_{DRP}) is most effective in detecting nonlinear effects?
2. Are these tests practically useful for detecting quadratic effects, interaction effects, or even both nonlinear effects simultaneously?
3. Is the performance of these tests influenced by the size of the nonlinear effects?
4. Is the performance of these tests influenced by additional nonnormality of the predictor variables?

METHOD

Two Monte Carlo studies were conducted with the aim of investigating the performance of standard and robust likelihood-based difference tests. Study 1 investigates the influence of nonlinear effect size and sample size on the performance of T_D, T_{DR}, and T_{DRP}. Study 2 investigates the influence of the distribution of the predictor variables on the performance of the different test statistics. For both studies the same population models were used for data generation and the same model difference tests were performed. The data were analyzed with the LMS approach under M*plus.* In the following, first, we introduce the model comparisons for difference testing; and second, we present the particular design for the two studies.

Model Comparisons

The same nested structural equation models were used throughout in Study 1 and Study 2. The first model (M_{LQI}) was a nonlinear model with two linear (L), one quadratic (Q), and one interaction (I) term:

$$\eta_1 = \alpha + \gamma_{11}\xi_1 + \gamma_{12}\xi_2 + \omega_{11}\xi_1^2 + \omega_{12}\xi_1\xi_2 + \zeta. \qquad (11)$$

where η_1 is the latent endogenous variable, ξ_1 and ξ_2 are the latent exogenous predictor variables, ζ is the disturbance term, and γ_{11} and γ_{12} are the linear effect parameters; ξ_1^2 is the latent quadratic term, ω_{11} is the quadratic effect parameter, $\xi_1\xi_2$ is the latent interaction term, ω_{12} is the interaction effect parameter and α is the intercept. The linear effect parameters were $\gamma_{11} = \gamma_{12} = .30$ in all conditions. The size of ω_{11} and ω_{12} was varied in the first study. Each latent variable (ξ_1, ξ_2, and η_1) had three indicator variables, all of them with factor loadings of .80, and the correlation between ξ_1 and ξ_2 was set to $\phi_{12} = .30$. The variances of the latent variables were all set to 1.0. The M_{LQI} model is displayed in Figure 1.

The second model (M_{LQ}) was a nonlinear model with two linear (L) and one quadratic effect (Q):

$$\eta_1 = \alpha + \gamma_{11}\xi_1 + \gamma_{12}\xi_2 + \omega_{11}\xi_1^2 + \zeta. \qquad (12)$$

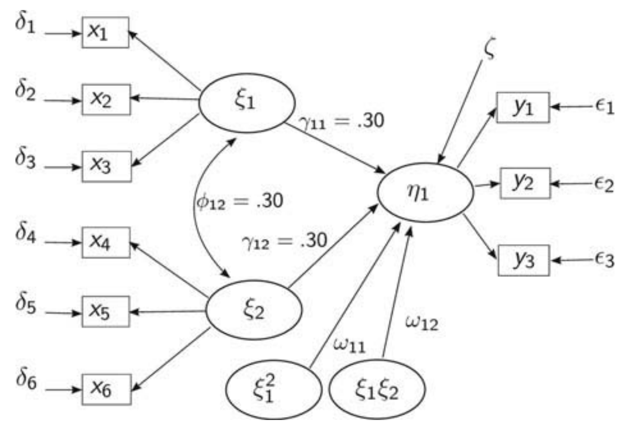

FIGURE 1 Nonlinear structural equation model M_{LQI} containing two linear effects (γ_{11}, γ_{12}), one quadratic effect (ω_{11}), and one interaction effect (ω_{12}). The reliabilities of the indicator variables were set to .64. The sizes of the nonlinear effects were varied in Study 1.

M_{LQ} is nested in M_{LQI}, because it only differs from M_{LQI} in setting the interaction effect ω_{12} to zero.

The third model (M_L) was a linear model with two linear effects (L):

$$\eta_1 = \gamma_{11}\xi_1 + \gamma_{12}\xi_2 + \zeta. \qquad (13)$$

M_L is nested in M_{LQ} and also in M_{LQI}, because both nonlinear effects (ω_{11}, ω_{12}) are set to zero.

The three nested models allow for the following model difference tests: (a) model with linear and quadratic effects compared to the linear model (M_{LQ} vs. M_L); (b) model with linear, quadratic, and interaction effects compared to the model with linear and quadratic effects (M_{LQI} vs. M_{LQ}); and (c) model with linear, quadratic, and interaction effects compared to the linear model (M_{LQI} vs. M_L).

Depending on the kind of analysis (Type I error or power analysis) the respective true population model differs. For power analysis M_1 is the true model and M_0 is the misspecified model. For Type I error analysis M_0 is the true model. For the different population models the model difference tests are listed with regard to the kind of analysis in Table 1.

TABLE 1
Population Models and the Model Difference Tests for Type I Error and Power Analyses

Population Model	*Type I Error*	*Power*
$\boldsymbol{M_{LQI}}$		$\boldsymbol{M_{LQI}}$ vs. M_{LQ} $\boldsymbol{M_{LQI}}$ vs. M_L
$\boldsymbol{M_{LQ}}$	M_{LQI} vs. $\boldsymbol{M_{LQ}}$	$\boldsymbol{M_{LQ}}$ vs. M_L
$\boldsymbol{M_L}$	M_{LQI} vs. $\boldsymbol{M_L}$ M_{LQ} vs. $\boldsymbol{M_L}$	

Note. M_L = linear model; M_{LQ} = model with linear and quadratic effects; M_{LQI} = model with linear, quadratic, and interaction effects. The respective population model is shown in bold type.

Study 1: Nonnormality Due to Nonlinear Terms

Study 1 was designed to examine the influence of nonnormality due to nonlinear terms on the test statistics. The performance of the likelihood-based T_D, $T_{DR,}$ and T_{DRP} was investigated. The sizes of the additional nonlinear effects in the less restrictive models (M_1) were varied. Overall, there were five effect size conditions for each model comparison: In model comparison (a) between M_{LQ} and M_L, the quadratic effect ω_{11} was set at .00, .15, .20, .25, and .30. A quadratic effect of .30 accounts for 18% unique variance in η_1. In model comparison (b) between M_{LQI} and M_{LQ}, the interaction effect ω_{12} was set at .00, .20, .25, .30, and .40. An interaction effect of .40 here accounts for 17.5% unique variance in η_1. As only the size of the additional effect of M_1 (the interaction effect) was varied, the quadratic effect was held constant, $\omega_{11} = .25$. In model comparison (c) between M_{LQI} and M_L, both nonlinear effects were varied correspondingly; the quadratic effect ω_{11} was set again at .00, .15, .20, .25, and .30; the interaction effect ω_{12} was set at .00, .20, .25, .30, and .40. The underlying population model differs, depending on whether the focus is on Type I error rate ($\omega_{ij} = .00$) or on power ($\omega_{ij} > .00$). The latent predictor variables ξ_1 and ξ_2 were normally distributed in the first study.

In all model comparisons the sample size N was set at 200, 400, and 800. In consideration that three model comparisons, five effect size conditions, and three sample size conditions were realized, altogether a design with $5 \cdot 3 \cdot 3 = 45$ difference tests was conducted in the first study.

The normally distributed data were generated with the PRELIS 2.8 software (Jöreskog & Sörbom, 2006). For each condition $R = 500$ replications were performed and the data were analyzed with M*plus* (Muthén & Muthén, 1998–2010) using the LMS method. For T_D, $T_{DR,}$ and $T_{DRP,}$ the Type I error rate was investigated. The proportion of data sets in which the M_1 models incorrectly fitted the data significantly better ($\alpha = 5\%$) than the true M_0 models was identified. Furthermore, the percentage of unusable negative difference values was calculated, and the power to detect nonlinear effects was investigated by identifying the proportion of data sets in which the M_1 models correctly fitted the data significantly better than the M_0 models.

Study 2: Nonnormality Due to Nonnormal Predictor Variables

Study II was designed to examine the influence of additional nonnormality due to nonnormally distributed predictor variables. Two distribution conditions were chosen:[3] (a) ξ_1 and ξ_2 were generated with a slight deviation from normality, where skewness of ξ_1 is $S_1 = 1$ and kurtosis is $K_1 = 5$, and skewness of ξ_2 is $S_2 = 1$ and kurtosis is $K_2 = 3$; (b) ξ_1 and ξ_2 were generated with a severe deviation from normality, where $S_1 = S_2 = 2$ and $K_1 = K_2 = 7$.

In addition to the distribution of the predictor variables, the sample size was varied as well, and $N = 200$ and $N = 400$ were tested. As the influence of the nonlinear effect sizes was already investigated in Study 1, here ω_{11} and ω_{12} were just varying depending on whether Type I error ($\omega_{ij} = .00$) or power ($\omega_{ij} > .00$) was analyzed. Thus, in each model comparison two effect sizes were realized: In comparison (a) between M_{LQ} and M_L, the quadratic effect ω_{11} was set at .00 and .25; in comparison (b) between M_{LQI} and M_{LQ}, the interaction effect ω_{12} was set at .00 and .30 and the quadratic effect was again held constant $\omega_{11} = .25$; and in comparison (c) between M_{LQI} and M_L, the nonlinear effects were varied correspondingly, ω_{11} was set at .00 and .25, and ω_{12} was set at .00 and .30. As the focus of this article is on the influence of nonlinearity on difference testing, we chose strong nonlinear effects for the less restrictive population models to gain more knowledge on the effects of noticeable nonlinearity. Overall, three model comparisons, two sample sizes, two effect sizes, and two distribution conditions were realized, resulting in a design with $3 \cdot 2 \cdot 2 \cdot 2 = 24$ difference tests for each testing procedure.

For nonnormal latent data generation, a two-step approach was performed. In the first step the data of the latent variables (latent predictor and error variables) were generated using the EQS program (Bentler, 2005) to ensure that the generated data met the assumed model parameters. In the second step the observable indicator variables were calculated from the previously generated latent variables, a procedure that is recommended for nonnormal data generation by Mattson (1997). For each condition 500 data sets were generated, and the performance of T_D, $T_{DR,}$ and T_{DRP} was investigated.

RESULTS

For the likelihood-based T_D, T_{DR}, and T_{DRP} we report (a) the Type I error rate, (b) the proportion of data sets with unusable negative difference values, and (c) the power to detect nonlinear effects.

Study 1: Nonnormality Due to Nonlinear Terms

The following results relate to the investigation of the influence of nonlinear effect size on the different likelihood-based difference statistics for models with normally distributed predictor variables.

[3]Note that it is not possible to choose arbitrary combinations of skewness and kurtosis. When the skewness is unequal to zero, the distribution is not symmetric anymore and a minimum value of kurtosis is required. Hence, the values of the kurtosis were chosen considering the possible values if a rather slight ($S = 1$) or a stronger ($S = 2$) skewness is given (cf. Werner, 2002).

TABLE 2
Type I Error Rates (%) for T_D, T_{DR}, and T_{DRP} as a Function of Sample Size (N)

Method	N		M_{LQ} vs. $\boldsymbol{M_L}$	M_{LQI} vs. $\boldsymbol{M_{LQ}}$	M_{LQI} vs. $\boldsymbol{M_L}$
		ω_{11}=	.00	.25	.00
		ω_{12}=	—	.00	.00
T_D	200		4.8	5.4	6.0
	400		4.4	4.2	4.0
	800		4.2	6.2	4.4
T_{DR}	200		7.4	7.2	8.0
	400		5.8	7.4	5.4
	800		5.0	8.0	6.2
T_{DRP}	200		2.0	2.6	1.6
	400		3.4	4.2	3.0
	800		3.2	6.0	4.0

Note. M_L = linear model; M_{LQ} = model with linear and quadratic effects; M_{LQI} = model with linear, quadratic, and interaction effects; ω_{11} = quadratic effect; ω_{12} = interaction effect. The respective population model is shown in bold type.

Type I error. In Table 2 the percentage of data sets is listed in which M_1 incorrectly fitted the data significantly better than M_0 ($\alpha = 5\%$). The Type I error rates in the simulation were close to the nominal 5% level for T_D, T_{DR}, and T_{DRP}.

T_D yielded a Type I error rate between 4% and 6.2%, the Type I error rate for T_{DR} ranged from 5% to 8%, and the Type I error rate for T_{DRP} ranged from 1.6% to 6%. The influence of sample size on Type I error was small. The occurrence of negative difference values was uncritical in these analyses (0% for T_D and T_{DRP}; 0%–0.4% for T_{DR}; these data sets were excluded from the calculation of Type I error rates).

Negative difference values. When the population model was M_1, T_{DR} frequently produced negative differences. This problem occurred notably when the nonlinear effect size of M_1 was large. T_D as well as T_{DRP} did not produce any negative difference values. Table 3 gives the percentage of data sets with negative T_{DR} values for the various model difference tests as a function of effect size and sample size. The percentage of negative T_{DR} values increased with increasing nonlinear effect size. When the additional nonlinear effects of M_1 were small (ω_{11} = .15, ω_{12} = .20) the problem of negative difference values was uncritical (0%–2.2%), whereas it increased considerably (23.0%–88.8%) when these effects were strong (ω_{11} = .30, ω_{12} = .40). Thus, under the most critical simulation condition, close to 90% of the T_{DR} values could not be used.

The influence of sample size on negative differences was less consistent. Under the condition of small nonlinear effects, an increasing sample size produced less negative difference values, whereas the risk of negative difference values increased with increasing sample size under the condition of strong nonlinear effects.

The type of model comparison influenced the risk of receiving negative difference values as well. The comparison between the quadratic model and the linear model M_{LQ} versus M_L produced fewer negative difference values than the comparison between the full nonlinear model and the linear model M_{LQI} versus M_L, with a maximum of 28.2% negative values in the former case and 88.8% negative values in the latter case.

Power. The power of T_D increased with increasing sample size and increasing effect size, whereas the power of T_{DR} and T_{DRP} decreased when nonlinear effect sizes were strong. Table 3 shows the results of the power analyses for the different model difference tests. The power of T_D, T_{DR}, and T_{DRP} is given as a function of effect size and of sample size. We report the percentage of data sets in which M_1 fitted the data significantly better than M_0.

The stronger the nonlinearity, the poorer T_{DR} and especially T_{DRP} performed, when compared to the well-performing T_D. The power of T_D increased with increasing nonlinear effect size and increasing sample size. This pattern could be observed across all model comparisons. Only when both the effect size and the sample size were small, a desirable power of 80% could not be reached under all conditions.

For T_{DR} the pattern was different: Initially the power also increased with increasing effect size, but when the nonlinear effects were strong the power decreased again, caused by data sets with negative difference values.[4] The influence of the sample size on the power of this test was similar to the influence of sample size on negative difference values shown earlier. When the effect sizes were small, the power increased with increasing sample size, and when the effect sizes were large, the power decreased with increasing sample size. In addition, the power of T_{DR} was lowest in the model comparison M_{LQI} versus M_L (*min* = 11.2%), when the nested linear model was severely misspecified.

For T_{DRP} an interesting pattern could be observed. When the sample size was large ($N \geq 400$) and the nonlinear effects were small, this test showed a substantial power. But even though this test showed no negative difference values at all, its power was low when the sample size was small, when the nonlinear effects were strong, or both. These problems were aggravated in model comparison M_{LQI} versus M_L, where the power of T_{DRP} was 0.0% when nonlinear effects were strong.

[4]Note that the data sets that produced negative difference values were not excluded from this analysis for two reasons: First, in some cases, a large amount of negative difference values occurred, so that exclusion would cause comparability problems between the different methods (e.g., 500 data sets of T_D and T_{DRP} vs. 56 data sets of T_{DR}); second, the difference values have to be comparably large to detect nonlinear effects reliably, therefore negative values cannot be ignored as they are causing the low power of T_{DR}.

TABLE 3
Power (%) of T_D, T_{DR}, and T_{DRP} and Negative T_{DR} Values (%, in brackets) as a Function of Nonlinear Effect Size and Sample Size (*N*)

			$\boldsymbol{M_{LQ}}$ vs. M_L				$\boldsymbol{M_{LQI}}$ vs. M_{LQ}				$\boldsymbol{M_{LQI}}$ vs. M_L			
		ω_{11}=	.15	.20	.25	.30	.25	.25	.25	.25	.15	.20	.25	.30
Method	*N*	ω_{12}=	—	—	—	—	.20	.25	.30	.40	.20	.25	.30	.40
T_D	200		72.0	90.0	98.4	99.8	61.2	79.6	93.0	100.0	96.6	100.0	100.0	100.0
	400		95.0	99.8	100.0	100.0	90.8	98.8	100.0	100.0	100.0	100.0	100.0	100.0
	800		99.8	100.0	100.0	100.0	99.6	100.0	100.0	100.0	100.0	100.0	100.0	100.0
T_{DR}	200		71.4	87.4	86.6	71.6	63.2	79.0	82.6	70.8	95.4	91.8	75.6	38.8
			(1.2)	(4.2)	(11.2)	(28.2)	(1.2)	(4.2)	(10.4)	(29.0)	(2.2)	(8.2)	(24.4)	(61.2)
	400		93.4	96.6	91.0	74.2	91.4	97.2	93.2	69.6	99.6	94.8	73.2	23.4
			(0.4)	(2.4)	(9.0)	(25.8)	(0.0)	(2.0)	(6.8)	(30.4)	(0.4)	(5.2)	(26.8)	(76.6)
	800		99.8	99.8	93.8	73.6	99.6	99.2	97.0	77.0	99.6	96.4	72.0	11.2
			(0.0)	(0.2)	(6.2)	(26.4)	(0.0)	(0.8)	(3.0)	(23.0)	(0.4)	(3.6)	(28.0)	(88.8)
T_{DRP}	200		31.4	41.8	38.0	22.6	45.0	64.0	78.6	64.8	50.8	26.6	6.0	0.0
	400		86.4	90.4	76.0	41.4	86.8	98.4	100.0	94.0	90.4	43.2	4.0	0.0
	800		99.8	100.0	92.4	51.2	99.6	100.0	100.0	99.8	98.4	51.6	1.4	0.0

Note. M_L = linear model; M_{LQ} = model with linear and quadratic effects; M_{LQI} = model with linear, quadratic, and interaction effects; ω_{11} = quadratic effect and ω_{12} = interaction effect. The respective population model is shown in bold type.

Study 2: Nonnormality Due to Nonnormal Predictor Variables

The following results relate to the investigation of the influence of nonnormally distributed predictor variables on the different statistics.

Type I error. For nonnormally distributed predictors, the Type I error rate for T_{DR} was greatly increased. Under the condition of severe deviation from normality, the Type I error rate for T_D was also inflated. Table 4 shows Type I error rates for T_D, T_{DR}, and T_{DRP} for the various model difference tests. The occurrence of negative differences was uncritical in this analysis (0% for T_D and T_{DRP}, 0% to 2.2% for T_{DR}; these data sets were excluded).

For T_D the Type I error rates ranged from 4% to 10.4%. They increased when there was a severe deviation from normality for the predictor variables. The Type I error rates for comparison M_{LQI} versus M_{LQ} were lower than for M_{LQI} versus M_L, which, in turn, were lower than for M_{LQ} versus M_L. The Type I error rates for T_{DR} increased even more steeply when the predictor variables were nonnormal. When the predictor variables showed a severe deviation from normality, the Type I error rate for T_{DR} was between 10.8% and 18.6%. The Type I error rates for T_{DRP} were then small (1.6%–7.8%) and often too conservative.

Negative difference values. In the power analyses T_{DR} frequently produced negative difference values when

TABLE 4
Type I Error Rates (%) for T_D, T_{DR}, and T_{DRP} as a Function of Deviation From Normality and of Sample Size (*N*)

		M_{LQ} vs. $\boldsymbol{M_L}$		M_{LQI} vs. $\boldsymbol{M_{LQ}}$		M_{LQI} vs. $\boldsymbol{M_L}$	
	ω_{11}=	.00		.25		.00	
	ω_{12}=	—		.00		.00	
Method	*N*	*Slightly Nonnormal*	*Severely Nonnormal*	*Slightly Nonnormal*	*Severely Nonnormal*	*Slightly Nonnormal*	*Severely Nonnormal*
T_D	200	7.4	10.2	6.0	5.8	6.8	8.0
	400	6.0	10.4	4.6	4.0	5.0	8.6
T_{DR}	200	14.0	17.2	11.2	10.8	16.4	18.6
	400	12.2	18.4	9.0	10.6	11.6	16.8
T_{DRP}	200	2.2	3.0	3.2	2.4	2.6	2.6
	400	5.4	7.8	3.0	4.0	2.0	5.2

Note. M_L = linear model; M_{LQ} = model with linear and quadratic effects; M_{LQI} = model with linear, quadratic, and interaction effects; ω_{11} = quadratic effect and ω_{12} = interaction effect. The respective population model is shown in bold type. Condition "slightly nonnormal": Skewness for ξ_1 and ξ_2 is $S_1 = S_2 = 1$; kurtosis is $K_1 = 5$ and $K_2 = 3$. Condition "severely nonnormal": $S_1 = S_2 = 2$ and $K_1 = K_2 = 7$.

TABLE 5
Power (%) of T_D, T_{DR}, and T_{DRP} and Negative T_{DR} Values (%, in Brackets) as a Function of Deviation From Normality and of Sample Size (N)

			$\boldsymbol{M_{LQ}}$ *vs.* M_L		$\boldsymbol{M_{LQI}}$ *vs.* M_{LQ}		$\boldsymbol{M_{LQI}}$ *vs.* M_L	
		ω_{11}=	.25		.25		.25	
		ω_{12}=	—		.30		.30	
Method	*N*		*Slightly Nonnormal*	*Severely Nonnormal*	*Slightly Nonnormal*	*Severely Nonnormal*	*Slightly Nonnormal*	*Severely Nonnormal*
T_D	200		99.8	99.8	83.4	84.4	100.0	100.0
	400		100.0	100.0	99.2	99.8	100.0	100.0
T_{DR}	200		16.0	19.6	65.0	53.0	11.2	12.8
			(83.4)	(80.2)	(22.6)	(34.2)	(88.8)	(87.2)
	400		5.6	6.2	73.2	52.4	3.4	3.2
			(94.4)	(93.8)	(26.0)	(47.2)	(96.6)	(96.8)
T_{DRP}	200		4.4	11.0	59.6	52.4	0.4	1.4
	400		3.6	10.6	94.6	94.4	0.0	0.2

Note. M_L = linear model; M_{LQ} = model with linear and quadratic effects; M_{LQI} = model with linear, quadratic, and interaction effects; ω_{11} = quadratic effect and ω_{12} = interaction effect. The respective population model is shown in bold type. Condition "slightly nonnormal": skewness for ξ_1 and ξ_2 is $S_1 = S_2 = 1$; kurtosis is $K_1 = 5$ and $K_2 = 3$. Condition "severely nonnormal": $S_1 = S_2 = 2$ and $K_1 = K_2 = 7$.

there was any deviation from normality. As in Study 1, T_D and T_{DRP} did not produce any negative differences. Table 5 contains the percentage of negative T_{DR} values for the model difference tests for the different distribution conditions with varying sample size.

T_{DR} has no benefit for nonlinear models, as the problem of negative differences, which was already observed in Study 1, was aggravated when the predictor variables showed any deviation from normality. When the latent predictor variables showed a slight deviation from normality, T_{DR} produced 22.6% to 96.6% negative values. When the deviation from normality was severe, the problem of negative differences became more serious in comparison M_{LQI} versus M_{LQ}. In comparisons M_{LQ} versus M_L and M_{LQI} versus M_L the percentage of negative differences was already high under the condition of slight deviation from normality and it remained constant under the condition of severe deviation from normality. The comparison M_{LQI} versus M_L yielded the most negative differences. Under the most critical simulation condition, close to 97% of the data sets showed negative T_{DR} values, and only 3% of the data sets produced usable results.

Power. The power of T_D exceeded 83% across all conditions, whereas the power of T_{DR} and T_{DRP} was very low when the predictor variables did not follow a normal distribution. Thus, only T_D was able to identify nonlinear effects correctly. Table 5 shows the power of T_D, T_{DR}, and T_{DRP} for the different simulation conditions and sample sizes.

T_D was robust against deviation from normality. In model comparison M_{LQI} versus M_{LQ}, the power was approximately 84% when the predictor variables were nonnormally distributed and sample size was small. Under the other conditions the power of T_D exceeded 99%.

The power of T_{DR} decreased when the predictor variables were nonnormal. This was due to negative difference values. Under the condition of severe deviation from normality, the power of T_{DR} ranged between 3.2% and 53.0%, depending on the type of model difference test. To compare, the power of T_{DR} was between 73.2% and 93.2% when the predictor variables were normally distributed (cf. Table 3). The power of T_{DR} was again most critical in model comparison M_{LQI} versus M_L, where T_{DR} identified nonlinear effects in merely 12.8% of the cases ($N = 200$) and ineffective with a power of 3.2% for $N = 400$ under the condition of severe nonnormality.

T_{DRP} reached a desired power of 80% in model comparison M_{LQI} versus M_{LQ} only, when sample size was at $N = 400$. The other model comparisons yielded adverse results: When the data showed slight or severe deviation from normality, the power of T_{DRP} was between 3.6% and 11.0% in comparison M_{LQ} versus M_L, and between 0% and 1.4% in comparison M_{LQI} versus M_L.

DISCUSSION AND OUTLOOK

In two Monte Carlo studies we examined the performance of likelihood-based model difference tests T_D, T_{DR}, and T_{DRP} for LMS in models with a quadratic or both a quadratic and an interaction effect. The robustness of T_D, T_{DR}, and T_{DRP} against different sources of nonnormality was investigated. In the first study the robustness against nonnormality due to the nonlinear model structure was investigated. The second study focused on the robustness against additional nonnormality due to nonnormally distributed predictors. Our results show that under the conditions examined, T_D is preferable to both T_{DR} and T_{DRP}.

T_D performed well under ideal conditions (Study 1). The nonlinear effects were reliably detected in most conditions. Thus, T_D appeared not only suited for the detection of interaction effects, but also for the detection of quadratic or multiple nonlinear effects simultaneously. However, it should be noted that, when the expected nonlinear effects were small, sample size needed to be large enough to ensure sufficient power.

T_D showed robustness against nonnormal latent predictor variables to a large extent across most conditions (Study 2). However, when the predictor variables had a severe deviation from normality and the test was executed at a 5% nominal Type I error level, the Monte Carlo Type I error rate was inflated up to 10% when a model with a quadratic effect was compared to a linear model. When predictor variables are nonnormally distributed, their third centered moments are nonzero, which implies that the nonlinear terms (quadratic and interaction terms) are correlated with the linear predictors. Although predictor variables in nonlinear SEM are generally centered, which reduces nonessential multicollinearity (cf. Aiken & West, 1991), multicollinearity due to nonnormally distributed predictors still remains. Such a correlation might lead to spurious interaction or quadratic effects and thus could inflate Type I error rates (cf. Klein, Schermelleh-Engel, Moosbrugger, & Kelava, 2009).

In the comparisons involving a model with both a quadratic and an interaction effect, an interesting pattern was observed: When the baseline model was quadratic and the comparison model included both nonlinear effects, the Type I error rates were close to their nominal alpha levels. However, when the baseline model was a linear model the Type I error rates were slightly inflated (up to 8%). To avoid spurious interaction effects, it is recommended to use the quadratic model for a baseline model (cf. Klein et al., 2009). The inclusion of the quadratic term (or even two quadratic terms) in the baseline model can considerably reduce the variance increment of the interaction term (cf. Ganzach, 1997; Lubinski & Humphreys, 1990; Klein et al., 2009). Therefore it is preferable to first test the model with both interaction and quadratic effects against a linear model and—when nonlinearity is detected in the data—the model with an interaction and a quadratic effect against a quadratic model.

When T_{DR} was used, several problems occurred during application. T_{DR} often produced unusable negative difference values, which has been already reported for linear SEM (Bryant & Satorra, 2012; Satorra & Bentler, 2010). This problem occurred especially when there were large nonlinear effects in the data, and the problem was aggravated when the latent predictors were nonnormally distributed. Then, the percentage of negative values exceeded 90% in some studies. The risk of Type I error also increased considerably when the predictor variables did not follow a normal distribution. T_{DR} appeared neither robust against nonnormality caused by a nonlinear model structure nor against nonnormality caused by nonnormal predictor variables. In the context of linear SEM, it has been noted that negative differences occur especially when the more restrictive model is highly incorrect (Satorra & Bentler, 2010). This is in accordance with the result that most of the negative differences occurred in our comparison when a nested misspecified linear baseline model was tested against a true population model with an interaction and a quadratic effect.

Using T_{DRP}, the occurrence of negative difference values could be avoided, but in turn this test statistic led to other problems. Under the condition of normally distributed latent predictor variables, T_{DRP} got smaller with increasing nonlinear effect size, and the power of T_{DRP} decreased. Small samples also affected the power of this test significantly, even when the nonlinear effect size was small. As for T_{DR}, the comparison between a true model with two nonlinear effects and a linear baseline model resulted in most unfavorable results. When the nonlinear effect sizes were very strong in this comparison, this test had virtually no statistical power, regardless of sample size. This means that T_{DRP} is vulnerable to highly misspecified nested models in combination with strong nonlinear effect sizes, just as is the $T_{DR.}$ Additional nonnormality due to the predictor variables also influenced this test significantly, as was the result in Study 2. See, for example, M_{LQ} versus M_L with $N = 400$ (cf. Table 3 and Table 5), where the power decreased from 76% to only 11% caused by a severe deviation from normality. Here again, there is no advantage for using T_{DRP} for detecting nonlinear effects, but rather a disadvantage compared to the uncorrected T_D.

Altogether, the standard test statistic T_D is the most practically useful method, albeit when further research is still necessary to evaluate the Type I error inflation under conditions of nonnormal predictor variables.

As we used the distribution analytic method LMS, the results presented in this article have limitations due to the chosen analysis approach. The nonnormality caused by the nonlinear terms is explicitly considered by distribution analytic methods and therefore LMS is a powerful method for analyzing nonlinear structural equation models. But LMS is just one possible approach to estimate nonlinear models. We did not investigate the PI approaches (e.g., Jöreskog & Yang, 1996; Marsh, Wen, & Hau, 2004; Moosbrugger et al., 2009), where products of the indicators are formed as a measurement model of the latent nonlinear term. In comparison to these PI approaches, the applied LMS approach does not need separate indicators for the latent nonlinear terms and it has less distributional assumptions. Hence, results of the presented simulation studies cannot be generalized to PI approaches. It is possible that PI approaches could lead to different results. Cham et al. (2012) investigated the robust SB chi-square difference test for PI approaches. The authors reported inflated Type I error rates and concluded that the robust difference test is unnecessary. Negative difference values occurred in the PI approaches as well, but the authors did not report their frequencies.

The PI approaches might benefit from the strictly positive difference test, which has not been examined for these approaches yet.

FUNDING

This research was supported by Grant No. SCHE 1412–1/1 from the German Research Foundation (DFG).

REFERENCES

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage.

Asparouhov, T., & Muthén, B. (2012). *Computing the strictly positive Satorra–Bentler chi-square test in M*plus (M*plus* Web Notes No. 12). Retrieved from http://www.statmodel.com/examples/webnotes/SB5.pdf

Bentler, P. M. (2005). EQS 6.1. for Windows [Computer software]. Encino, CA: Multivariate Software.

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

Bryant, F. B., & Satorra, A. (2012). Principles and practice of scaled difference chi-square testing. *Structural Equation Modeling*, *19*, 372–398.

Cham, H., West, S. G., Ma, Y., & Aiken, L. S. (2012). Estimating latent variable interactions with nonnormal observed data: A comparison of four approaches. *Multivariate Behavioral Research*, *47*, 840–876.

Chou, C. P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology*, *44*, 347–357.

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

Dimitruk, P., Schermelleh-Engel, K., Kelava, A., & Moosbrugger, H. (2007). Challenges in nonlinear structural equation modeling. *Methodology*, *3*, 100–114.

Ganzach, Y. (1997). Misleading interaction and curvilinear terms. *Psychological Methods*, *2*, 235–247.

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

Hughes, J. N., & Kwok, O. (2006). Classroom engagement mediates the effect of teacher–student support on elementary students' peer acceptance: A prospective analysis. *Journal of School Psychology*, *43*, 465–480.

Jöreskog, K., & Sörbom, D. (2006). PRELIS (Version 2.8) [Computer software]. Chicago, IL: Scientific Software International.

Jöreskog, K. G., & Yang, F. (1996). Nonlinear structural equation models: The Kenny–Judd model with interaction effects. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 57–87). Mahwah, NJ: Erlbaum.

Kelava, A., Werner, C. S., Schermelleh-Engel, K., Moosbrugger, H., Zapf, D., Ma, Y., Cham, H., Aiken, L. S., & West, S. G. (2011). Advanced nonlinear latent variable modeling: Distribution analytic LMS and QML estimators of interaction and quadratic effects. *Structural Equation Modeling*, *18*, 465–491.

Klein, A. G. (2000). *Moderatormodelle: Verfahren zur Analyse von Moderatoreffekten in Strukturgleichungsmodellen* [Moderator models: Method for the analysis of moderator effects in structural equation models]. Hamburg, Germany: Dr. Kovac.

Klein, A. G., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, *65*, 457–474.

Klein, A. G., & Muthén, B. O. (2007). Quasi maximum likelihood estimation of structural equation models with multiple interaction and quadratic effects. *Multivariate Behavioral Research*, *42*, 647–673.

Klein, A. G., & Schermelleh-Engel, K. (2010). Introduction of a new measure for detecting poor fit due to omitted nonlinear terms in SEM. *AStA Advances in Statistical Analysis*, *94*, 157–166.

Klein, A. G., Schermelleh-Engel, K., Moosbrugger, H., & Kelava, A. (2009). Spurious interaction effects. In T. Teo & M. S. Khine (Eds.), *Structural equation modeling in educational research: Concepts and applications* (pp. 13–28). Amsterdam, The Netherlands: Sense.

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

Lubinski, D., & Humphreys, L. G. (1990). Assessing spurious "moderator effects": Illustrated substantively with the hypothesized ("synergistic") relation between spatial and mathematical ability. *Psychological Bulletin*, *107*, 385–393.

Marsh, W. H., Wen, Z., & Hau, K.-T. (2004). Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction. *Psychological Methods*, *9*, 275–300.

Marsh, W. H., Wen, Z., Nagengast, B., & Hau, K.-T. (2012). Structural equation models of latent interactions. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 436–458). New York, NY: Guilford.

Marshall, G. N., Miles, J. N. V., & Stewart, S. H. (2010). Anxiety sensitivity and PTSD symptom severity are reciprocally related: Evidence from longitudinal study of physical trauma survivors. *Journal of Abnormal Psychology*, *1*, 143–150.

Mattson, S. (1997). How to generate non-normal data for simulation of structural equation models. *Multivariate Behavioral Research*, *32*, 355–373.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156–166.

Moosbrugger, H., Schermelleh-Engel, K., Kelava, A., & Klein, A. G. (2009). Testing multiple nonlinear effects in structural equation modeling: A comparison of alternative estimation approaches. In T. Teo & M. S. Khine (Eds.), *Structural equation modeling in educational research: Concepts and applications* (pp. 103–135). Rotterdam, The Netherlands: Sense.

Moosbrugger, H., Schermelleh-Engel, K., & Klein, A. (1997). Methodological problems of estimating latent interaction effects. *Methods of Psychological Research Online*, *2*, 95–111.

Muthén, B. O., & Muthén, L. K. (2006). *Chi-square difference testing using the Satorra–Bentler scaled chi-square* [Weblog post]. Retrieved from http://www.statmodel.com/chidiff.shtml

Muthén, L. K., & Muthén, B. O. (1998–2010). M*plus* (Version 6.1) [Computer software]. Los Angeles, CA: Authors.

Muthén, L. K., & Muthén, B. O. (2010). *M*plus *user's guide*. Los Angeles, CA: Authors.

Satorra, A. (1990). Robustness issues in structural equation modeling: A review of recent developments. *Quality & Quantity*, *24*, 367–386.

Satorra, A. (1992). Asymptotic robust inferences in the analysis of mean and covariance structures. *Sociological Methodology*, *22*, 249–278.

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In D. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis: A festschrift for Heinz Neudecker* (pp. 233–247). Dordrecht, The Netherlands: Kluwer Academic.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*, 507–514.

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, *75*, 243–248.

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika*, *50*, 253–264.

Werner, C. S. (2002). *Robustheit der LMS-Methode zur Analyse latenter Moderatormodelle bei systematischer Verletzung der Verteilungsvoraussetzungen* [Robustness of LMS for the analysis of latent moderator models to systematic violations of distributional assumptions] (Unpublished diploma thesis). Goethe University, Frankfurt Germany.

Willoughby, M., Cadigan, J., Burchinal, M., & Skinner, D. (2008). An evaluation of the psychometric properties of and criterion validity of the Religious Social Support Scale. *Journal for the Scientific Study of Religion*, *47*, 147–159.