Normality Assumption in Latent Interaction Models

Sirio Lonati[1], Mikko Rönkkö[2], John Antonakis[3]

[1] NEOMA Business School

[2] Jyväskylä University School of Business and Economics, University of Jyväskylä

[3] Department of Organizational Behavior, University of Lausanne

Author note

Sirio Lonati https://orcid.org/0000-0003-0216-2313

Mikko Rönkkö https://orcid.org/0000-0001-7988-7609

John Antonakis https://orcid.org/0000-0001-8811-5117

We thank Jeffrey Edwards, Charles Efferson, Alberto Holly, Christian Zehnder and the participants of the OB brownbag at the University of Lausanne for their comments and for helpful discussions. John Antonakis also acknowledges the FNS Grant number IZ70Z0_131326 "The Influence of Entrepreneurial Leadership and Social Capital on Resource Assembly and Firm Performance in Small- and Medium-Sized Firms in East Africa". Mikko Rönkkö acknowledges the Academy of Finland grant number 311309. We acknowledge the computational resources provided by the Aalto Science-IT project.

A previous, early version of this work was presented at the Academy of Management Annual Meeting, 2020, held online (https://doi.org/10.5465/AMBPP.2020.18911abstract).

The simulation code used in this study is available at https://osf.io/bsx23/?view_only=92ccaa76a26740578f984c8348fe2723.

Correspondence concerning this article should be addressed to Sirio Lonati, NEOMA Business School, 59 rue Pierre Taittinger, 51100 Reims, France. E-mail: Sirio.lonati@neoma-bs.fr

Abstract

Latent Moderated Structural Equations (LMS) is one of the most common techniques for estimating interaction effects involving latent variables (i.e., `XWITH` command in Mplus). However, empirical applications of LMS often overlook that this estimation technique assumes normally distributed variables and that violations of this assumption may lead to seriously biased parameter estimates. Against this backdrop, we study the robustness of LMS to different shapes and sources of non-normality and examine whether various statistical tests can help researchers detect such distributional misspecifications. In four simulations, we show that LMS can be severely biased when the latent predictors or the structural disturbances are non-normal. On the contrary, LMS is unaffected by non-normality originating from measurement errors. As a result, testing for the multivariate normality of observed indicators of the latent predictors can lead to erroneous conclusions, flagging distributional misspecification in perfectly unbiased LMS results and failing to reject seriously biased results. To solve this issue, we introduce a novel Hausman-type specification test to assess the distributional assumptions of LMS and demonstrate its performance.

Keywords: Interactions, Moderation, Latent variables, LMS, Hausman, Specification test, Instrumental variables, `XWITH`

Testing the normality assumption in latent interaction models

Moderating effects—also called interaction effects—are ubiquitous in the social sciences. However, the correct estimation of interaction effects presents several critical challenges, including the issue of measurement error (e.g., Cortina et al., 2021; Dawson, 2014; Dimitruk et al., 2007; Kenny & Judd, 1984). Fortunately, several techniques for modeling measurement error in moderated models exist (Kelava & Brandt, 2022). Among these methods, Latent Moderated Structural Equation Modeling is perhaps the most popular (LMS, Klein & Moosbrugger, 2000). The success of LMS is easily explained. If the model is correctly specified, LMS produces consistent (i.e., the parameter estimates converge to the true value as sample size increases) and efficient (i.e., precise) estimates. LMS is also easy to apply, adding to its popularity (e.g., the `XWITH` command in Mplus or the `nlsem` package in R, Umbach et al., 2017)[1].

Whereas LMS is straightforward to apply, understanding whether it produces trustworthy results is not as simple. Consider the following example using the Program for International Student Assessment (PISA) 2006 dataset inspired from Kelava et al. (2014). We use LMS and another technique, EXT (explained later in the article), to study if students' career aspiration in science depends on their enjoyment of science, their academic self-concept in science, and the interaction between the two. Both LMS and EXT are consistent when their assumptions hold and should produce very similar results in large samples such as the PISA data. Nevertheless, this time, LMS and EXT return entirely different results. LMS suggests a negative but not significant interaction effect (-.02, SE = .02, p = .344), whereas EXT suggests a positive and strongly

[1] To get a rough idea of how widespread LMS is, it is worth reporting that a simple search on Google Scholar on 19.11.2023 of the keywords *("LMS" OR "XWITH") AND "MPLUS"* returns 2,550 results.

statistically significant (.05, SE = .02, p = .009) effect. Why are these results so different? Which technique should one trust?

When two consistent techniques produce vastly different results in large samples, this strongly indicates that the assumptions of at least one of the techniques do not hold. Thus, we need to focus on the assumption that LMS makes and EXT does not: Normality of *latent* variables. This normality assumption is different from the assumption of normally distributed *observed* data (i.e., indicators of latent variables) made in linear latent variable models (Klein & Moosbrugger, 2000). It is also a much stronger assumption. In linear models, violations of normality lead to biased standard errors and fit statistics, but they leave the parameter estimates unaffected, at least in large samples (e.g., Curran et al., 1996; Satorra & Bentler, 2001; Yuan et al., 2005). When using LMS, non-normal latent variables not only lead to incorrect inferential tests but can also lead to biased parameter estimates (e.g., Aytürk et al., 2020; Brandt et al., 2014; Cham et al., 2012; Klein & Muthén, 2007). As such, failing to acknowledge potential biases related to non-normality in empirical applications that rely on LMS or just assuming that Satorra-Bentler corrections or similar fixes will solve them risks engendering misleading conclusions (e.g., Alessandri et al., 2021; Belogolovsky et al., 2016; Fasbender & Gerpott, 2023; McDermott et al., 2020; Sun et al., 2021; Taggart et al., 2019). Thus, statistical procedures and methodological guidelines for testing the normality assumption of LMS are sorely needed.

In this paper, we tackle this broad topic by focusing on two types of diagnostic tools: multivariate normality tests for the observed indicators (e.g., Doornik & Hansen, 2008; Mardia, 1970) and a novel Hausman-type specification test (cf. Hausman, 1978). Multivariate normality tests rely on a simple intuition. If the observed indicators of the latent predictors are significantly non-normal, chances are that latent predictors are also non-normal, thus causing biased LMS estimates. The specification test rests on a different logic, which, to the best of our knowledge,

was first proposed informally by Aguirre-Urreta et al. (2020). Specifically, it involves comparing LMS (i.e., a consistent and efficient estimator under normality) with another estimator that may be less efficient but is known to be consistent under more relaxed assumptions. If the two estimators produce very similar point estimates in large samples, this result means that the underlying population data are likely normal and LMS can be trusted; if they differ, normality is likely violated, suggesting that LMS is biased and inconsistent and should not be used to estimate the data at hand.

We contribute to the literature by formalizing the novel specification test and comparing its performance against traditional multivariate normality tests in four Monte Carlo simulations. We show how LMS and alternative estimators perform under a large set of non-normality shapes (e.g., skewed or kurtotic variables) and types (i.e., non-normal latent predictors, structural disturbance, or measurement errors of the observed indicators). Overall, our results suggest that the multivariate normality tests and the specification tests are complementary. We, thus, urge all future applications of LMS to be more mindful of its normality assumption, always to report both types of tests in research articles, and to prefer a consistent estimator whenever LMS' distributional assumptions are violated.

Measurement error in interaction models: Problems and solutions

Before turning to the core topic of our paper—the normality assumption—we briefly explain some key ideas related to measurement error in interaction models. We then present common corrective procedures, differentiating between estimators that assume normality (LMS) and estimators that do not.

Measurement error in moderated models: Latent interaction

Social scientists often use imprecise measures from questionnaires, surveys, or archival data. In these cases, measurement error—the difference between an observed measure and its

latent, unobserved value—can lead to untrustworthy estimates (Antonakis et al., 2010; Blake & Gangestad, 2020; Cole & Preacher, 2014; Cortina et al., 2021; Schmidt & Hunter, 1996). Measurement error in the predictors is problematic because it causes the well-known attenuation bias in bivariate regression, making it harder for researchers to detect effects. However, when multiple independent variables are measured with error, the bias's magnitude and direction are difficult to predict (Wooldridge, 2002, Chapter 9).

Researchers in psychology and neighboring disciplines often solve the measurement error issue using latent variable models (cf. Kline, 2015). This approach is straightforward if a model contains only linear terms, yet it becomes complex if it includes latent interactions. Consider the simplest possible latent interaction model shown in Figure 1 (see equations 1-6 in Klein & Moosbrugger, 2000 for details). This model has an observed dependent variable, y, and two latent predictors, ξ_1 and ξ_2. The latent variables ξ_1 and ξ_2 represent unobservable concepts, which are measured with three observed indicators (e.g., questionnaire items; x_1, x_2, and x_3 for ξ_1; z_1, z_2, and z_3 for ξ_2). δ_1 to δ_6 represent measurement errors, and ϵ is the disturbance of the structural equation. The measurement errors and the disturbance are assumed to be uncorrelated with the latent predictors and each other.

Estimating the model in Figure 1 is complicated by the fact that the latent variables ξ_1 and ξ_2 do not have observed values that could be multiplied together to produce the interaction term, $\xi_1\xi_2$, as one would do when estimating interactions between observed variables. The literature has, thus, proposed three broad categories of estimation techniques to "produce" $\xi_1\xi_2$ and estimate its effect (for a review, see, e.g., Kelava & Brandt, 2022): Distribution-analytic approaches, product indicator approaches, and observed variables approaches. We next review one commonly used estimator from each category: LMS, the extended unconstrained product

indicators (EXT, Marsh et al., 2004), and model-implied instrumental variables (MIIV, Bollen, 1996; Bollen & Paxton, 1998)[2]. These techniques differ in their technical implementation, ease of use, efficiency, and—crucially for this article—distributional assumptions (Brandt et al., 2020).

[Figure 1 about here]

Latent Moderated Structural Equations

Among the distribution analytic approaches, LMS is the most widely adopted. The mathematical explanation of LMS may appear complex (see Klein & Moosbrugger, 2000), but the intuition behind it is simple. Assume that the latent predictors, the measurement errors, and the structural disturbance are normally distributed. Because the product of two normal latent predictors is non-normally distributed, the dependent variable of an interaction model is also non-normal. LMS estimation relies on this fact: The larger the interaction effect, the more non-normal the dependent variable should be. Formally, LMS does this by approximating the non-normal multivariate density of the indicators by a finite mixture of conditionally normal distributions produced by a numerical integral over the distribution of the latent predictor variables (Klein & Moosbrugger, 2000).

When latent predictors, measurement errors, and structural disturbance are normally distributed, LMS is a maximum likelihood estimator and is consistent, asymptotically efficient, and asymptotically normal (Klein & Moosbrugger, 2000). Simulation results confirm these properties in finite samples (Brandt et al., 2014). However, LMS can perform poorly when its distributional assumptions fail (Kelava & Nagengast, 2012). The distribution of the latent

[2] Other approaches include, for example, the Quasi-Maximum Likelihood method (Klein & Muthén, 2007), the Method of Moments of Wall and Amemiya (2003), the method of Mooijaart and Bentler (2010), and the nonlinear structural equation mixture modeling approach (Kelava et al., 2014; Umbach et al., 2017). We do not cover these approaches because they are not widely used.

predictors is particularly critical. Generally, the stronger the non-normality, the larger the bias in the LMS estimates (Kelava et al., 2014; Lodder et al., 2019). The bias can also increase—rather than decrease—in larger sample sizes (Aytürk et al., 2020; Cham et al., 2012).

Extended Unconstrained Product Indicator approach

The model in Figure 1 can also be estimated by adding a new latent variable representing the interaction term (i.e., $\xi_1\xi_2$) and using the products of the indicators (e.g., x_1z_1) as indicators for this new variable. When two variables are multiplied together, the error term of the product generally correlates with the error terms of the original variables (e.g., the product indicators x_1z_1 and x_1z_2 have the linear indicator x_1 in common, Foldnes & Hagtvet, 2014). This issue can be addressed in many ways by selecting which product indicators are used, centering the original variables or the product indicators, and freeing and constraining the error covariances of the product indicators. Consequently, many variants of this "product indicator approach" have been proposed since the early 1980s (Kenny & Judd, 1984). Recent literature recommends the "extended unconstrained product indicator" approach (EXT, Kelava & Brandt, 2009). This straightforward strategy freely estimates all parameters in latent interaction models (Marsh et al., 2004).

EXT is typically estimated using the same maximum likelihood estimator used for models containing only linear terms. This estimator is consistent and asymptotically normal when the data are multivariate normal (equation 2, see Yuan et al., 2005) and remains consistent irrespective of the distribution of the data (see, e.g., Browne & Shapiro, 1988; White, 1982; Yuan et al., 2005). However, both theory (see, e.g., eq. 25, Wall & Amemiya, 2001) and simulation work (e.g., Curran et al., 1996) show that standard errors and fit statistics are not trustworthy with non-normal variables. This property can represent a problem for EXT because the product indicators are never normally distributed. Fortunately, simulation studies show that the bias of

EXT is small and, although EXT underestimates standard errors (Cham et al., 2012; Kelava et al., 2014; Marsh et al., 2004), this problem can be solved by using robust standard errors. However, robust standard errors are valid only asymptotically and may still be slightly underestimated in small samples (Brandt et al., 2014).

Model Implied Instrumental Variables

The MIIV procedure of Bollen (1996) is an entirely different way to estimate latent interaction models. MIIV recasts the latent variable model as an observed variable one by expressing each latent variable as a function of one of its indicators (latent-to-observed transformation). For instance, the two latent variables of Figure 1 can be expressed as $\xi_1 = x_1 - \delta_1$ and $\xi_2 = z_1 - \delta_4$ (see equations 7-11, Bollen & Paxton, 1998). The transformation introduces a correlation between the predictors and the structural disturbance term and must be estimated with an instrumental variable estimator. In the MIIV estimator, the other indicators of $\xi_1$and ξ_2 (i.e., x_2, x_3 and z_2, z_3) can serve as instruments, because each indicator is correlated with the latent variable, yet should be uncorrelated with the structural disturbance (Bollen et al., 2022). When the interaction term $x_1 z_1$ is included in the model, the products of the instruments (i.e., $x_2 z_2, x_2 z_3, x_3 z_2$, and $x_3 z_3$) are used as additional instruments (see, e.g., Brandt et al., 2020).

The MIIV estimator is typically estimated with two-stage least squares (Bollen, 1996) or generalized method of moments (Bollen et al., 2014). As such, MIIV does not require any distributional assumptions for consistency and asymptotic normality (Bollen, 1996; Bollen & Paxton, 1998). However, MIIV is inefficient and biased in small samples (see Chapter 4, Angrist & Pischke, 2009; Klein & Moosbrugger, 2000; Moulder & Algina, 2002); perhaps for this reason, there has not been much simulation work on MIIV estimation of interaction models.

The normality assumption

The normality assumption is a major difference between LMS and the alternative estimators (EXT and MIIV). Whereas LMS has been proven to be consistent only when latent predictors, measurement errors, and structural disturbance(s) are normally distributed, EXT and MIIV do not require any specific distributional assumption to achieve the same result. This raises a deceptively simple question: Why would researchers ever want to use LMS instead of the safer alternatives that make fewer assumptions?

Normality assumption: Convenience or conviction?

Under normality, LMS is both statistically more appealing and more convenient to apply than the alternatives. First, it is more efficient (Brandt et al., 2014; Kelava et al., 2014). This means that researchers have a higher chance of correctly finding a significant interaction effect using LMS than other estimators. Second, LMS calculates correct standard errors under normality without the corrections to fit statistics and standard errors that EXT requires. Third, LMS does not suffer from the small sample bias problem of MIIV. Last, LMS can be effortlessly evoked with canned routines (e.g., the `XWITH` command in Mplus or the `nlsem` package in R, Umbach et al., 2017), and it does not require constructing complex product indicators, choosing which products to use, and deciding how to specify them in the model.

Unfortunately, the normality assumption is not always credible. Indeed, observed data in the social sciences are often non-normal (Cain et al., 2017). Such non-normality can originate from three sources: the latent predictor variables, the structural disturbances, or the measurement errors. Non-normality in the latent predictors emerges whenever researchers focus on a variable that would be skewed or kurtotic even if it was perfectly measured. For instance, non-normal latent predictors can emerge whenever a researcher studies a rare ability (see, e.g., Micceri, 1989), a low base-rate phenomenon (e.g., abusive supervision, Fischer et al., 2021), or a power-

law distributed variable, like individual performance (i.e., many low-performers and a handful of star performers are usually observed, see Aguinis & O'Boyle Jr, 2014).

Yet, even when latent predictors are (approximately) normally distributed (e.g., intelligence, Antonakis et al., 2017), non-normality might creep in from measurement errors. Non-normal measurement errors emerge by construction whenever researchers measure a (possibly normal) continuous latent factor with a Likert scale containing only categorical response options. In this case, respondents cannot precisely indicate their true score of the latent variable, ending up indicating the categorical option lying closer to their true latent value, thereby causing a non-normal measurement error[3]. Skewed measurement errors are also possible. For instance, the wording of questionnaire items might systematically shape responses, causing some ceiling effects (cf. Schwarz, 1999). Similarly, respondents' reactions to sensitive or socially desirable questions might cause skewed responses, wherein most respondents misrepresent their true answers and fake either complete agreement or disagreement on a question (see Katz & Katz, 2010; Millimet & Parmeter, 2022).

Finally, a non-normal structural disturbance also violates LMS assumptions and can occur even if latent predictors and measurement errors are normally distributed. The disturbance represents the sum of all unmeasured causes of the outcome, which can be many. The normality assumption of the disturbance is often justified by arguing that the sum of many independent causes is generally normal regardless of the distribution of the individual causes (and, thus, the

[3] To better understand how the use of categorical indicators leads to non-normal errors, the reader can run the following simple R code:

```
set.seed(123)
latent.var <- rnorm(10000)                    # normal latent variable
observed.indicator <- round(latent.var)       # observed indicator
meas.error <- latent.var - observed.indicator # measurement error
hist(meas.error)
```

Central Limit Theorem applies, Zeckhauser & Thompson, 1970). However, there can sometimes be just a handful of omitted causes. For instance, if a researcher's model already explains most of the outcome variance, assuming many unobserved causes might be unwarranted. It is also possible that one unobserved cause correlates with other ones, leading to non-independent omitted causes. The justification based on the Central Limit Theorem also breaks apart if one of the unobserved causes is substantially more important than others. Finally, measurement errors of observed dependent variables are comprised in the disturbance and can cause non-normality, as explained in the previous paragraph.

Testing the normality assumption: Multivariate normality of the observed indicators

A key question we address is this: How do we test for the inherently *unobservable* normality assumption of LMS? The first way is to examine the distribution of the observed indicators of the latent predictor variables (Brandt et al., 2014; Klein & Moosbrugger, 2000). Because observed indicators are functions of the latent variable and measurement error (e.g., $x_1 = \xi_1 + \delta_1$ in Figure 1), non-normality of an indicator means that either the latent variable, ξ_1, or the measurement error, δ_1, or both are non-normally distributed. Thus, researchers can inspect the normality of the indicators using graphical aids or multivariate normality tests (e.g., Mardia, 1970) and use this information to determine if LMS estimates can be trusted.

Yet, inspecting the distribution of the observed indicators has some limits. First, multivariate normality tests cannot determine whether non-normality originates from measurement error or latent predictors. This is a critical limitation because prior research suggests that non-normal latent predictors cause significant problems for LMS, but the same is

not necessarily the case with non-normal errors[4]. Second, multivariate normality tests cannot be applied to the dependent variable and, thus, cannot detect non-normality that originates from the structural disturbance. The issue is that the dependent variable is expected to be non-normal whenever an interaction effect exists (i.e., the product between two normal variables is never normally distributed). As such, testing the distribution of the outcome is meaningless because a non-normal outcome can either signal the presence of a non-normal disturbance or of an interaction effect. Third, multivariate normality tests of observed data are not tests for latent variable distributional misspecifications—and hence for bias—of LMS. As such, they risk flagging trivial deviations from normality that might lead to negligible bias in LMS (e.g., Klein & Moosbrugger, 2000; Lodder et al., 2019).

Testing the normality assumption: Hausman-type specification test

Because of the limitations of multivariate normality tests, we need a diagnostic tool that does not target the observed indicators. Moreover, we are not interested in violations of normality *per se* but rather in non-normality that is large enough to bias the LMS estimator meaningfully. Thus, we follow the intuition of Aguirre-Urreta et al. (2020) and propose a novel Hausman-type specification test to test the distributional assumptions of LMS.

Hausman's (1978) specification test is a general test of the consistency of an estimator that applies to a wide range of empirical problems (e.g., Aït-Sahalia & Xiu, 2019; Creel, 2004; Fair & Parke, 1980; Hahn & Hausman, 2002; White, 1980). The idea of the test is the following. Consider two different ways of estimating the same parameter (e.g., a latent interaction term)

[4] To the best of our knowledge, no simulation has tested the performance of LMS with non-normal continuous measurement errors, yet several papers have shown that categorical indicators (which can be thought of as a way to introduce a specific type of non-normal measurement error) lead to relatively little bias in LMS *per se* (Aytürk et al., 2020, 2021; Lodder et al., 2019).

referred to as $\hat{\beta}_E$ and $\hat{\beta}_C$. Both estimators are consistent but $\hat{\beta}_C$ makes fewer assumptions than $\hat{\beta}_E$. When its stronger assumptions hold, $\hat{\beta}_E$ is more efficient than $\hat{\beta}_C$. For instance, $\hat{\beta}_E$ might be efficient and consistent only under normality (e.g., LMS). In contrast, $\hat{\beta}_C$ might be consistent under all distributional conditions but is less efficient (e.g., EXT, MIIV). In large samples, the difference between two consistent estimators should be small. The Hausman test considers this difference. If the difference is small enough to be attributed to chance, we can conclude that $\hat{\beta}_E$ is consistent. If the difference is large, we must conclude that $\hat{\beta}_E$ is inconsistent.

More formally, the Hausman test tests the null hypothesis that $\hat{\beta}_E$ is consistent. The test assumes that $\hat{\beta}_E$ is efficient, that $\hat{\beta}_C$ is consistent but inefficient, and that both estimators are normal in large samples under the null hypothesis. The test statistic is the ratio of the squared difference between the estimators and the variance of this difference:

$$H = \frac{\left(\hat{\beta}_C - \hat{\beta}_E\right)^2}{Var\left(\hat{\beta}_C - \hat{\beta}_E\right)} \quad (1)$$

This ratio follows a χ^2 distribution with 1 degree of freedom under the null hypothesis (Hausman, 1978, Theorem 2.1). The key challenge in calculating the test statistic is estimating $Var\left(\hat{\beta}_C - \hat{\beta}_E\right)$. Hausman's (1978) main contribution was proving that $Var\left(\hat{\beta}_C - \hat{\beta}_E\right) = Var\left(\hat{\beta}_C\right) - Var\left(\hat{\beta}_E\right)$. This produces a simple test statistic (Hausman, 1978, Lemma 2.1):

$$H = \frac{\left(\hat{\beta}_C - \hat{\beta}_E\right)^2}{Var\left(\hat{\beta}_C\right) - Var\left(\hat{\beta}_E\right)} \quad (2)$$

A key challenge in the Hausman test is that Equation (2) is not always defined. Because $Var\left(\hat{\beta}_C\right)$ and $Var\left(\hat{\beta}_E\right)$ are estimated, the (estimated) variance of the efficient estimator is sometimes larger than the (estimated) variance of the consistent estimator even when all the assumptions of the test hold. When this occurs, the χ^2 value is negative and has no *p*-value. This

issue does not invalidate the Hausman test in large samples (as it becomes irrelevant asymptotically), but it poses significant challenges in practical applications. There are ways to address the issue with various closed-form estimators (e.g., Baum et al., 2003), but they do not apply to maximum likelihood estimates that use numerical optimization.

A general solution to the negative *H* problem is to use another statistic that approximates the Hausman test statistic but is always positive. These approaches are sometimes called "Robust Hausman Tests" in the literature (Cameron & Trivedi, 2005, p. 273). One way to define such a statistic for a single coefficient is (henceforth labeled "robust Hausman")

$$H_R = \frac{\left(\hat{\beta}_C - \hat{\beta}_E\right)^2}{Var\left(\hat{\beta}_C\right)} \tag{3}$$

The H_R statistic does not follow any known probability distribution because the variance of the numerator depends on $\hat{\beta}_E$, which is not included in the denominator. However, H_R can be a reasonable proxy of H because it is always smaller than H[5]. This means that if H_R is larger than the $\chi^2(1)$ critical value used in the significance test, then the null hypothesis would also be surely rejected based on H.

Another solution to the negative Hausman statistic is to use a lower bound for the test statistic—that is, the lowest possible value the Hausman statistic can have—under maximally conservative assumptions (Cameron & Trivedi, 2005, p. 273):

$$H_{LB} = \frac{\left(\hat{\beta}_C - \hat{\beta}_E\right)^2}{Var\left(\hat{\beta}_C\right) + Var\left(\hat{\beta}_E\right)} \tag{4}$$

[5] Proving that H_R is always smaller than or equal to H is straightforward: $Var(\hat{\beta}_C)$ is equal to or larger than $Var\left(\hat{\beta}_C\right) - Var\left(\hat{\beta}_E\right)$ under the null hypothesis, because $Var\left(\hat{\beta}_E\right)$ is always a positive value.

Like H_R, the H_{LB} statistic is always positive. H_{LB} is also always smaller than H under the null hypothesis, assuming that $Cov(\hat{\beta}_C, \hat{\beta}_E) > 0$, because the denominator of Equation (4) **Error! Reference source not found.**is always greater than the denominator of Equation (1).

In sum, H_R and H_{LB} are conservative versions of the standard Hausman test. These statistics are guaranteed to be always positive and smaller than H. However, both H_R and H_{LB} are less powerful than the standard Hausman test (Cameron & Trivedi, 2005, pp. 273-274). This property should guarantee Type I error rates below these tests' nominal levels (e.g., less than 5%). That is, both tests have a low probability of incorrectly flagging an unbiased LMS estimate as misspecified. The downside is that they risk flagging biased LMS estimates as correctly specified because the tests are less powerful than the standard Hausman test (see Figure 2).

[Figure 2 about here]

Testing the normality assumption: Simulation evidence

How well do the multivariate normality tests of the observed indicators and the Hausman specification tests of LMS against EXT or MIIV detect distributional misspecifications in LMS in finite samples? In this section, we answer this question with four Monte Carlo simulations where we manipulate both the shape of non-normality and its origin (i.e., latent predictors, measurement errors, or structural disturbance). Table 1 summarizes the simulation setup and main results.

In Simulation 1, we manipulate latent predictors' distributions while keeping both measurement errors and the structural disturbance normally distributed. This scenario should lead to large biases in LMS and substantial departures from normality in the observed indicators. Thus, we expect the performance of multivariate normality tests to be excellent. Because latent predictors' non-normality should leave the parameters estimated by EXT and MIIV unaffected,

we expect the specification tests to also perform well. However, all specification tests should lack power compared to the multivariate normality tests.

In Simulation 2, we use normal latent predictors and structural disturbance, and manipulate the observed indicators' measurement error distribution. We expect the multivariate normality tests to flag any possible distributional misspecifications in LMS, as non-normal measurement errors will substantially affect observed indicators' distributions. However, we do not have a clear expectation about LMS' performance. Whereas non-normal measurement errors formally represent a distributional violation for LMS (Klein & Moosbrugger, 2000), prior simulation evidence suggests that non-normal measurement errors have little biasing effect on LMS. Thus, we expect the specification test to rarely—if ever—flag LMS models as distributionally misspecified.

In Simulation 3, we focus on a scenario that mimics the use of rating scales in a questionnaire (cf. Rhemtulla et al., 2012). In this simulation, we use non-normal latent predictors and normal structural disturbance (as in Simulation 1) but introduce also non-normal errors by collapsing the indicators into a small number of categories. We expect the multivariate normality tests to flag most data of this kind as problematic because categorical indicators are, by definition, non-normal. However, we also expect LMS to be biased mainly by the non-normal distribution of the latent predictors and not by the categorical indicators, at least if the number of categories of the indicators is large enough (Aytürk et al., 2020). Thus, we expect the multivariate normality tests to flag as misspecified some LMS estimates that might not be particularly biased and that the specification test might not reject.

Finally, in Simulation 4, we let non-normality enter only from the structural disturbance. In this scenario, we have no clear expectations about the performance of LMS because violations of such distributional assumptions have never been studied, to the best of our knowledge. As a

result, we have no clear prediction about the performance of the specification tests. We anticipate, however, that the multivariate normality tests should be useless in this scenario because the misspecification affects only the dependent variable. As we previously discussed, the outcome's distribution is uninformative (and, hence, not tested), because it cannot be expected to be normal even in a correctly specified interaction model.

[Table 1 about here]

Overview of the simulations

Because all four simulations share the same basic structure, we set up the study as one large fractional factorial design. Each of the four simulations was full-factorial and nested in this larger design. The R code for the simulation and data are available online at https://osf.io/bsx23/?view_only=92ccaa76S26740578f984c8348fe2723.

Data generation

We use Figure 1 to set up a basic latent interaction model[6]. The coefficients of ξ_1 and ξ_2 (i.e., γ_1 and γ_2) were set to 1, the variance of the structural disturbance was set at 3.2, and the nonlinear effect of $\xi_1\xi_2$ (i.e., γ_3) was set to .50 to produce effect sizes that are comparable with prior simulations (Brandt et al., 2020; Marsh et al., 2004; Moulder & Algina, 2002) or to 0 for a null interaction effect. The latent variables correlate .30. The error variances of x_1-x_3 and z_1-z_3 were set to either .86 or 1.14, corresponding to a coefficient alpha of roughly .80 and .70, respectively. The sample size was varied at 100, 200, 350, 650, and 1,000.

All simulations used the same distributions but varied how the non-normal distributions were applied. In the baseline condition, all latent predictor variables, structural disturbance, and

[6] We use an observed dependent variable for simplicity and because measurement error in the dependent variable does not cause bias (Wooldridge, 2002, Chapter 9). As a robustness check, we re-run simulation 1 with a latent dependent variable, finding qualitatively identical bias figures (unreported results).

indicator error terms were normal. We used several non-normal distributions for the other conditions. We had four χ^2 distributions, with 1, 2, 4, or 8 degrees of freedom, ranging from severe to low skewness (cf. Kelava et al., 2014). We also had a *t* distribution with 4 degrees of freedom to model data with relatively large kurtosis (i.e., platykurtic data) and a uniform distribution, which models data with negative excess kurtosis (i.e., leptokurtic data). The distributions were scaled to have a mean of 0 and a standard deviation of 1 in the population. We used the copula method by Mair et al. (2012) to generate correlated non-normal latent variables because of the issues identified in the more commonly used Vale and Maurelli (1983) method (Foldnes & Grønneberg, 2015; Olvera Astivia & Zumbo, 2015). The indicator error terms (i.e., the measurement errors δ_1-δ_6 in Figure 1) and the structural disturbance (ϵ in Figure 1) are independent of other variables and were, thus, generated from univariate distributions. We used 1'000 replications per condition.

Estimators

For each family of estimators considered, we employed several variants. First, we used two options for LMS: The traditional LMS and the reliability-corrected single-indicator LMS (SI-LMS), recently introduced in prominent organizational behavior journals (Cheung & Lau, 2017; Su et al., 2019). SI-LMS differs from the traditional LMS because it combines the indicators of each latent predictor as a single parcel whose error variances are constrained like in errors-in-variables models (Culpepper, 2012). The LMS-SI proponents argue that its theoretical properties should be close to those of LMS, yet violations of the distributional assumptions might have less severe consequences (Su et al., 2019). We also estimated LMS with robust standard errors as a robustness check. All LMS estimates were calculated with Mplus using the MplusAutomation package (Hallquist & Wiley, 2018).

Second, we used two variants of EXT, one with matched product indicators (EXT-MATCHED) and one with all product indicators included (EXT-ALL). EXT-MATHCED was set up following Marsh et al. (2004) using only three non-redundant product indicators (i.e., x_1z_1, x_2z_2, and x_3z_3) for the non-linear term. EXT-ALL was set up following Foldnes and Hagtvet (2014), using all potential product indicators (i.e., $x_1z_1, x_2z_2, x_3z_3, x_1z_2, x_1z_3, x_2z_1, x_2z_3, x_3z_1, x_3z_2$) and freeing all error covariances among product indicators sharing an indicator (e.g., x_1z_1 and x_1z_2). The indicators were centered before generating the product indicators (Marsh et al., 2006; Wall & Amemiya, 2001). The EXT estimates were calculated with lavaan (Rosseel, 2012) using robust standard errors (indicated with the "-R" label) and, as a robustness check, also with conventional ones.

Third, we used two variants of MIIV. MIIV-ALL was set up using all possible product indicators as instruments. Thus, there are 2 instruments per latent variable (i.e., x_2, z_2, x_3, and z_3) and 4 instruments for the interaction term (i.e., x_2z_2, x_2z_3, x_3z_2, and x_3z_3). MIIV-PARCEL combined the original instruments as three parcels (see Rönkkö et al., 2020). Parceling the instruments is an unexplored strategy, which might be helpful in small samples and follows the recommendation by Bollen et al. (2007) to use only a few instruments. The MIIV estimates were calculated with a two-stage least squares estimator implemented with the AER package (Kleiber & Zeileis, 2009).

Finally, as a benchmark, we also estimated a naïve observed-variable regression model via ordinary least squares (OLS) using scale means and their interactions as predictors.

Non-convergent solutions and outliers

Before analyzing the simulation data, we inspected solutions flagged as nonconvergent or where no coefficients or standard errors were produced. After ensuring that these were caused by

a non-convergent estimator and not a programming error, we discarded the entire replication (i.e., all estimators) because a non-convergent result signals a replication that could also be challenging for the other estimators. Then, following Boomsma (2013), we used the box-plot method to identify outliers in the interaction effect estimates. Proceeding with one estimator-design combination at a time, we flagged all estimates that were more than three interquartile ranges below the first quartile or above the third quartile as outliers. We then deleted all replications that contained at least one outlier estimate, as in the case of non-convergent replications.

Performance of the estimators

We analyze the bias of the estimators using absolute bias (i.e., the average estimated interaction effect over successful replications minus the true value of the interaction effect). To assess their precision (i.e., efficiency), we calculated the mean standard deviation of the estimated interaction effect across all successful replications in each design cell. We then compared this statistic against the mean standard error to assess the bias of the standard errors (see, e.g., Morris et al., 2019; see also Supplementary Material 1).

Multivariate normality and specification tests

We tested the multivariate normality of the observed indicators using the Mardia's multivariate kurtosis and skewness tests (Mardia, 1970) and the more recent omnibus test of Doornik and Hansen (2008), all calculated with the MVN package (Korkmaz et al., 2014)[7]. We also recorded the H_R, H_{LB}, and the standard Hausman statistic and their p values.

[7] During the study, we noted that the result of the Doornik-Hansen test computed by the MVN package depended on the means of the variables. After verifying that this behavior was caused by a programming error in the package, we reported the error to the package maintainer. Because we did not hear back from him, we forked the package and published a corrected version at https://github.com/mronkko/MVN. We also simulated the Royston

Simulation 1: Manipulating latent predictors' distributions

In Simulation 1, we manipulated the distribution of the latent predictor variables using 11 different configurations. We used the three symmetric distributions (normal, *t*, uniform) and four χ^2 distributions. Each χ^2 distribution was used twice by modeling a scenario where both predictors were skewed in the same direction (right-skewed) and a scenario where they were skewed in different directions. We used all five sample sizes and varied the indicators' reliability and whether an interaction existed in the population, producing an 11×5×2×2 = 220 full-factorial design. Structural disturbance and measurement errors were standard normal in all conditions.

Results for Simulation 1

Figure 3 shows the overview of the simulation results across latent predictors' distributions, and Figure 4 and Figure 5 show the results also by sample size. Although we show the results for all estimators in Figure 3, we discuss only on the best-performing estimator of each type: LMS with normal standard errors, EXT-ALL-R with robust standard errors, and MIIV-PARCEL[8]. Complete results are reported in Supplementary Material 1.

Outliers and non-convergent solutions. Overall, there were a few convergence problems with LMS (.07%) and EXT-ALL-R (.71%). MIIV-PARCEL always converged, being a closed-form estimator. Outliers were also rare: LMS (.20 %), EXT-ALL-R (.76 %), and MIIV-PARCEL (.78%). Most outliers emerged in the smallest sample condition, which was the most challenging one for all estimators.

(Royston, 1992) and the Henze-Zirkler (Henze & Zirkler, 1990) multivariate normality tests. Their performance was qualitatively similar to the Mardia tests and the Doornik-Hansen test. Given that these tests are also more rarely used by applied researchers, we just report their results in the Supplementary Material.

[8] We focused on: (a) LMS with normal standard errors because the use of robust standard errors did not make a difference and because the performance of LMS-SI, while being promising and generally better behaved than LMS at the smallest sample size, led to no substantive differences in terms of bias under non-normality; (b) EXT-ALL-R because non-robust standard errors are generally underestimated and because EXT-MATCHED-R is slightly more biased; (c) MIIV-PARCEL, which performs similarly to MIIV-ALL but tends to be less biased.

Bias when the latent interaction effect is present. LMS shows virtually no bias in the normal condition. However, when the latent variables are highly skewed in the same direction, LMS is strongly biased (i.e., absolute bias around .40). The bias decreases with decreasing skewness (i.e. when the degrees of freedom of the χ^2 distribution increases) and when the latent variables are skewed in opposite directions (see Figure 4). To our understanding, this result is new to the literature. When both predictors are skewed in the same direction, the outcome contains a lot of "excess skew" compared to the case of normal predictors. Our intuition is that because interactions of correlated and centered variables also produce skewed outcome distributions, the excess skew is "used" by LMS to infer the existence of an interaction effect, causing bias. When the predictors are skewed in different directions, there is much less excess skew and consequently a smaller positive bias in LMS estimates (i.e., the opposite skews "cancel out"). With symmetric distributions, LMS shows sizable bias in the *t*-distribution condition but virtually no bias in the uniform condition. We believe this pattern emerges because interactions also produce heavy-tailed distributions (large kurtosis). Thus, the *t* distribution with its large kurtosis throws off the LMS estimator in a way that the uniform distribution with its negative excess kurtosis does not.

In contrast to LMS, the performance of both MIIV-PARCEL and EXT-ALL-R is virtually unaffected by non-normally distributed predictors. Also, indicator reliability has little effect on their bias. In contrast, when LMS produces biased results because of non-normal predictors, the bias becomes visibly larger in low-reliability conditions (see Figure S1). Moreover, the effects of sample size on LMS' bias are minor compared to the effects of the latent variables' distributions (see Figure S4). Finally, the linear terms estimated by LMS are also slightly biased under non-normality (see Figure S10).

Bias when the latent interaction effect is null. The same pattern of results mostly holds when there is no interaction effect in the population, as shown in Figure 5. One important difference is that LMS is now unbiased when the latent variables are skewed in opposite directions. Moreover, when LMS produces biased results, the bias tends to be less pronounced than in the conditions where the effect existed in the population. We attribute both results to the previously mentioned LMS's sensitivity to skewed outcomes. However, the bias LMS can still be unacceptably high in some conditions; in these cases, EXT and MIIV are unbiased and should be preferred. Finally, the effect of indicators' reliability and sample size is qualitatively similar to the conditions where the latent interaction effect is present (see Figures S1 and S7).

Efficiency and accuracy of the standard errors. Figure 3 shows that LMS is the most efficient estimator with a clear margin when its assumptions hold, especially in the smallest sample sizes (see Figure S5). LMS remains more efficient than the alternative unbiased estimators also in some non-normal conditions. However, LMS tends to lose its efficiency advantage when skewness increases, becoming even less efficient than EXT-ALL-R in the most skewed condition. All these patterns are qualitatively similar, whether the interaction effect exists or not (see Figure S8).

Concerning the other estimators, MIIV-PARCEL is generally the least efficient estimator. Moreover, MIIV-PARCEL also underestimates standard errors in the most skewed conditions. EXT-ALL-R is generally more efficient and its estimated standard errors tend to be closer to the Monte Carlo standard deviations. EXT approaches tend, however, to overestimate the Monte Carlo standard deviation at the smallest sample sizes and especially in uniform and normal conditions, confirming known results (see Figure S6-S9).

Multivariate normality and specification tests. The Mardia and the Doornik-Hansen tests perform very well, flagging biased LMS models even in the smallest sample sizes (Figure 4

and Table S5). The Mardia skewness test is especially powerful. On the flip side, the multivariate normality tests also detect instances of non-normality that are inconsequential for LMS, especially when the interaction effect does not exist (see Figure 5 and Table S6). This is because non-normal latent predictors invariably cause non-normal observed indicators, yet not all shapes and degrees of non-normality are equally problematic for LMS.

Turning our attention to the specification test, we find a large number of negative Hausman statistics. Averaging across conditions, almost 64% and 45% of H values computed using EXT-ALL and MIIV-PARCEL as consistent estimators were negative. The negative *H* issue was more common in the most skewed condition (85.78% and 72.25% negative values) than in the normal one (27.61% and 6.17% negative values) because LMS is no longer efficient in these conditions. Although it may be tempting to interpret a negative *H* as evidence against the null hypothesis of no distributional misspecification, the false positive rates (i.e., when all latent variables are normally distributed) are too high for this decision rule to be of much practical use. As such, we do not discuss the traditional Hausman test any further and focus on the robust alternatives.

The robust Hausman specification test has adequate power in mid-to-large sample sizes and for severely skewed latent variables, especially when using EXT-ALL-R as the consistent estimator (Figure 4 and Table S5). The test loses power when distributions are less skewed, or the predictors are skewed to different directions because LMS is less biased in these conditions. The H_R test is more powerful than the "lower bound" Hausman test, H_{LB}, by a clear margin in all conditions (in line with Cameron & Trivedi, 2005, p. 273). Both specification tests show a low Type I error rate (less than 5%) when latent predictors are normally distributed, regardless of which estimator is used as the consistent one. In the case of a null interaction effect, the power of the specification tests is visibly lower (Figure 5 and Table S6). This low statistical power stems

from the fact that, even though the distributional assumptions of LMS are violated, they affect the LMS estimates less than in the case where the interaction effect exists. Finally, the specification test performs particularly well in the low-reliability conditions when the latent interaction effect is present, most likely because LMS's performance degrades particularly in these cases (see Figure S11-S14).

[Figure 3, Figure 4, and Figure 5 about here]

Brief discussion of Simulation 1

Simulation 1 demonstrates that the specification tests can flag important distributional misspecifications at mid-to-large sample sizes. However, the multivariate normality tests of the observed indicators can accomplish the same objective with substantially more statistical power. These tests are so powerful that using them to detect misspecifications in LMS might come at a cost. Whereas all non-normal distributions formally violate LMS's assumptions, some violations are inconsequential, leading researchers to select possibly less efficient estimators over LMS even when LMS would show minimal bias.

Simulation 2: Manipulating measurement errors' distributions

In Simulation 2, we use normal latent predictors and introduce non-normality through indicators' measurement errors. To save on computation costs, we use a subset of the conditions from Simulation 1. Instead of using all χ^2 distributions, we used only the most skewed one and, thus, focused on three distributions ($\chi^2(1)$, $t(4)$, and uniform). We did not include the normal distribution because the "all normal" condition was already reported as a part of Simulation 1. The sample sizes and reliability conditions were the same as in the first simulation, producing a 3×5×2 = 30 full factorial design.

Results

For brevity, we again report only results for the estimators that performed best in Simulation 1 and do not report figures for standard errors and Monte Carlo standard deviations (complete results are in Supplementary Material 2).

Outliers and non-convergent solutions. As in Simulation 1, we first dropped all non-convergent replications (.09% for LMS and 1.68% for EXT-ALL-R) and all replications including outliers (.05% for LMS, .80% for EXT-ALL-R, .66% for MIIV-PARCEL).

Bias. Figure 6 shows that measurement errors' distributions have little to no effect on the bias of any of the estimators, including LMS. This result is relatively unsurprising given that LMS estimates the interaction effect from the distribution of the dependent variable, which is unaffected by the distribution of the errors of the predictor variables' indicators.

Efficiency and accuracy of the standard errors. LMS is unaffected by measurement errors' distributions and generally remains the most efficient estimator, also estimating standard errors rather accurately (see Figures S20 and S23). In contrast, EXT-ALL-R (and, to a lower extent, also MIIV-PARCEL) tend to overestimate standard errors, particularly at low sample sizes (see Figures S21 and S24).

Multivariate normality and specification tests. The multivariate normality tests can detect all non-normal measurement errors throughout most conditions. Nevertheless, this high power is not very useful because non-normal errors pose little problems for LMS. To reiterate, this behavior is expected because the normality tests inform us of the distribution of the indicators but do not consider the bias of LMS estimates. In contrast, the specification tests perform better. Because of the low bias of LMS, the H_R and H_{LB} specification tests have a virtually 0 probability of flagging LMS estimates as misspecified, irrespective of whether EXT-ALL-R or MIIV-PARCEL is used as the consistent estimator.

[Figure 6 about here]

Brief discussion of Simulation 2

Simulation 2 clarifies that violations of distributional assumptions concerning measurement errors do not lead to meaningful bias in LMS. In this case, the multivariate normality tests flag unbiased LMS estimates as misspecified. In contrast, the Hausman-type specification test performs better, keeping the Type I error rate at a minimum. Taken together, the results thus far indicate that the multivariate normality tests and the specification tests have different strengths and weaknesses. The multivariate normality tests are very powerful in detecting problematic conditions for LMS, but using them as the sole diagnostic would lead to rejecting perfectly acceptable LMS results. In contrast, the specification tests are less powerful and outright underpowered in the smallest sample sizes but rarely lead to rejecting acceptable LMS results. This pattern suggests that the two classes of tests are complementary.

Simulation 3: Non-normal measurement error due to categorical observed indicators

In Simulation 3, we focused on categorical indicators (e.g., rating scales in a survey). We manipulated the distribution of the latent predictor variables using both correct (normal) and a subset of the misspecified ($\chi^2(1)$, $t(4)$, uniform) distribution conditions from the first simulation. Following Simulations 1 and 2, we used all five sample sizes and the two reliability conditions. Additionally, we varied the number of categories between 2 and 7 indicators, producing a 4×5×2×6 = 240 full factorial design. In all conditions, the structural disturbance was normally distributed. We generated categorical data by generating continuous indicators (exactly as in Simulations 1 and 2), cutting these into the desired number of categories, and finally rescaling the data to have the same population means and variances as the continuous data. The cutoffs were determined by estimating each continuous indicator's 1% and 99% cumulative probability and cutting the resulting interval into equally spaced parts.

Results

For brevity, we again report only the main results in the text (complete results are in Supplementary Material 3).

Outliers and non-converging simulations. We dropped all replications that did not converge (1.02% for LMS, 3.37% for EXT-ALL-R, and .001% for MIIV-PARCEL[9]) and all replications including outliers (.31% for LMS, 1.48% for EXT-ALL-R, 1.44% for MIIV-PARCEL). Compared to the previous simulations, these figures are larger. For instance, the non-convergence rates were as high as 50% for the LMS estimator in the most demanding conditions (e.g., binary indicators, skewed latent predictors, low reliability, and smallest sample size)[10]. We inspected a subset of the nonconvergent replications (using LMS-SI, which suffered especially from non-convergence issues) to understand the cause of the problem. Even increasing the number of integration points by ten times or using the population values as starting values did not help. Based on these diagnostics, we conclude that binary indicators present a challenging case for the LMS estimation algorithm. As such, the simulation results obtained when indicators have only two categories should be interpreted with much caution. In conditions with more categories, however, the non-convergence issue is much less severe.

---

[9] The MIIV estimator is closed form and, thus, does not have convergence issues. In this simulation, we encountered a few cases where binary indicators and the smallest sample size produced perfect collinearity between predictors or instruments. In this case, the estimates do not exist, and none of the estimation techniques provided results.

[10] LMS-SI had the most convergence problems (in the 90% range) in the two-indicator, low-reliability, and skewed latent predictors conditions. Whereas we present the results only for LMS, EXT-ALL-R, and MIIV-PARCEL in the main text, we always dropped non-convergent replications based on all estimators because the figures of Simulation 1 include all estimators. We use the same dropping rule throughout the study for simplicity and to maintain comparability between the simulations. Nevertheless—and because of the concern that the high non-convergence rates of LMS-SI might be driving the results in Simulation 3—we generated an alternative set of results where we excluded replications only based on LMS, EXT-ALL-R, and MIIV-PARCEL. These results are very similar to the figures presented in the article and they are available at https://osf.io/bsx23/?view_only=92ccaa76S26740578f984c8348fe2723.

Bias. Figure 7 shows that all estimators performed very similarly to Simulation 1, though the bias is systematically larger due to the lower information content of categorical indicators. LMS is severely biased when the latent predictors are skewed but is less affected by the kurtotic ones[11]. All estimators show some bias in the *t* distribution but are unaffected by the uniform one. A higher number of categories tends to decrease the bias for all estimators (see also Aytürk et al., 2020). This result aligns with a prior suggestion that items with enough categories (e.g., five) can usually be considered continuous (Rhemtulla et al., 2012)[12]. As in the previous simulations, when LMS is biased, the magnitude of the bias is strongly affected by the indicator reliability (see Figure S30-S31).

Efficiency and accuracy of the standard errors. The behavior of all estimators is similar to Simulation 1 but suffers when indicators have few categories. LMS is usually the most efficient estimator, except again in the most skewed condition. EXT-ALL-R and MIIV-PARCEL have large Monte Carlo standard deviations, especially in the *t*-distribution condition and when the indicators have only a handful of categories (see Figure S32-S33). All estimators tend to overestimate standard errors, but this problem diminishes from three or four categories onwards. When indicators are binary, all estimators substantially overestimate standard errors, especially in low-reliability conditions (see Figure S34-S35).

---

[11] Figure 7 shows a smaller bias for LMS in the two-category case compared to the three-category case. This result is an artifact due to how non-convergent results were dropped. When we drop results only based on LMS, EXT-ALL-R, and MIIV-PARCEL, LMS shows slightly more bias for the two-category case than the three-category case (see Footnote 10).

[12] Mplus implements categorical variable estimation using polychoric correlations. We also considered using this estimator but did not, because it would produce results that are scaled differently from the continuous variable estimators. Whereas the scaling issue can be solved by using standardized estimates (Rhemtulla et al., 2012), this procedure presents an additional problem because it is not clear how the LMS results should be standardized.

Multivariate normality and specification tests. The multivariate normality tests perform similarly to Simulation 1 but with some exceptions. Most notably, the Mardia multivariate kurtosis test and the Doornik-Hansen test flag most of the normal latent variable cases as non-normal. Although categorical indicators are indeed always non-normal, they rarely bias LMS per se, as our results show. Even with seven categories, these tests suggest discarding perfectly unbiased LMS estimates more than 50% of the time. In addition, there is some loss of statistical power compared to Simulation 1. Specifically, the Mardia kurtosis test flags now fewer issues when the latent variables are χ^2 distributed and the number of categories is more than 3. Likewise, the Mardia skewness test detects fewer issues when latent variables are *t* or uniformly distributed.

The specification tests also perform qualitatively similarly to Simulation 1. This is expected because LMS remains biased with non-normal latent variables even if indicators are categorical. However, H_R loses power because the standard errors of EXT and MIIV are overestimated and because these estimators can be slightly biased by the categorical indicators. The same applies to the H_{LB} statistic, which is particularly underpowered. Finally, like in Simulation 1, the specification tests perform better with low-reliability indicators, probably because of the LMS' bias, which is more considerable in this condition (see Figure S37-38).

[Figure 7 about here]

Brief discussion of Simulation 3

Simulation 3 shows that non-normality that originates from using categorical indicators to measure continuous latent variables does not, *per se*, lead to a large bias in LMS. Rather, categorical indicators worsen the bias caused by non-normal latent predictors. Categorical indicators pose challenges for both the multivariate normality tests and the specification tests.

The multivariate normality tests are the most powerful alternative for detecting problems when they exist but suffer from the problem of flagging inconsequential non-normality as problematic. Categorical indicators pose little-to-no problems for LMS as long as the other assumptions hold, but are flagged as non-normal and thus potentially problematic by the normality test. The Hausman-type specification test (H_R) performs visibly better in normal and uniform conditions, correctly suggesting that few LMS models are biased. However, the Hausman-type test has less power to detect problematic conditions, and its power further decreases with categorical indicators compared to continuous ones, making it useful only with extreme violations of normality.

Simulation 4: Manipulating structural disturbance

In Simulation 4, we again generate normal latent predictors and measurement errors but simulate non-normal structural disturbance. To save on computation costs, we focused again only on three distributions ($\chi^2(1)$, $t(4)$, and uniform). The sample size and reliability conditions were the same as in the other simulations. Altogether, this produced a 3×5×2 = 30 full factorial design.

Results for Simulation 4

For brevity, we again report only the main results in the paper (complete results are in Supplementary Material 4).

Outliers and non-convergent solutions. We first dropped from the analysis all non-convergent replications (1.76% for LMS and 1.62% for EXT-ALL-R) and all replications with outliers (1.46% for LMS, .93% for EXT-ALL-R, .67 for MIIV-PARCEL).

Bias. Figure 8 shows that all estimators are virtually unaffected by non-normal structural disturbances, with one exception. LMS suffers from a major bias when the disturbance is $\chi^2(1)$ distributed and the indicator reliability is low. The bias also counterintuitively increases with the

sample size. The magnitude of bias is surprising and deserves some explanation. We, thus, plotted the distribution of these estimates in Figure 9, which shows them to be strongly bimodal. The issue affected both LMS and LMS-SI. We investigated the latter because it was slightly less affected by the problem. To rule out the possibility of a programming error, we ran a few models by hand directly in Mplus. We also adjusted the number of integration points and starting values, which had little effect on the results. We then regressed the estimates against the sample characteristics and found that estimates close to the second mode depended on the skewness of the dependent variable in the sample. Plotting the estimate over skewness revealed a clear cutoff of skewness (about 1.7 for N = 1’000 and in the low-reliability condition), after which the secondary mode appears. Increasing the sample size to 5’000 decreased the secondary peak but did not eliminate it. These results and the datasets are available on the previously linked OSF page.

This unexpected behavior of LMS is cause for concern and should be studied further in future research. Here, we provide an initial and brief explanation of why and when this problem occurs. The product of two correlated latent variables has a skewed distribution. When a case has a dependent variable value far in the tail, it should also have explanatory variables’ values in the tails of their distributions. In the high-reliability conditions, we have more information on the explanatory variables; consequently, this second condition is easier to evaluate. When reliability decreases, it becomes increasingly difficult to differentiate whether the skewness of the dependent variable is due to the interaction or the skewed disturbance. As a result, the interaction effect becomes nearly empirically underidentified. In this scenario, the optimizer sees a flat region in the likelihood and tends to converge on the edge of this region (note: this pattern was visible in the Mplus iteration log). In other words, the estimation algorithm either attributes most of the (excess) skewness to the error term, producing an estimate in the primary mode, or

attributes most of the (excess) skewness to the interaction term, producing an estimate in the secondary mode.

Efficiency and accuracy of the standard errors. In the kurtotic and normal distributions, the behavior of the various estimators is similar to Simulation 1, with LMS being the most efficient. With highly skewed structural disturbances, however, LMS behaves poorly, being the least efficient estimator and severely underestimating the standard errors, especially in the low-reliability condition (see Figure S43; note, however, that interpreting the Monte Carlo standard deviation of LMS in the skewed condition is complex, due to the large bias exhibited by LMS in this condition). Concerning the other estimators, EXT-ALL-R shows an upward bias of the estimated standard errors (especially in small samples), whereas MIIV-PARCEL is mostly unaffected by the structural disturbance distribution (see Figures S43 and S46).

Multivariate normality and specification tests. The multivariate normality tests cannot provide any information about the distribution of the structural disturbance because these tests are —by construction—not applied to the dependent variable. In contrast, the specification tests perform better. Indeed, the H_R specification had virtually 100% chance of detecting extremely biased estimates in the secondary mode. The reason why Figure 8 (i.e., $\chi^2(1)$ – low reliability panel) shows lower power levels is just an artifact of the bimodal distribution of the LMS bias in this condition. Extreme bias (i.e., secondary mode) emerges in a subset of replications (between about 20% and 75% depending on the sample size) and the power levels in the figure reflect this fact (see Figure S50).

[Figure 9 and Figure 8 about here]

Brief discussion of Simulation 4

Simulation 4 portrays a complex and perplexing picture of LMS. Non-normal structural disturbances are mostly harmless, at least regarding bias. However, they can lead to rare—though

catastrophic—bias if the disturbance is severely skewed and the reliability of the indicators is relatively low. Such an extreme scenario might be rare in empirical studies. Nevertheless, it poses a major challenge for LMS and multivariate normality tests of the observed indicators, which cannot detect such misspecification. However, the Hausman-type test performs much better, flagging all instances of extreme LMS bias.

Empirical examples

We now present two empirical examples to show how to diagnose violations of normality in practice. The code to replicate the first empirical example, based on publicly available data, is available on the previously linked OSF page. For brevity, we present only the interaction models. The confirmatory factor analysis results, model fit indices, and diagnostics are presented in Supplementary Material 5. We emphasize that these examples are just illustrations and put aside other aspects of correct model specification (e.g., the presence of unobserved confounders). As such, the reader should refrain from drawing any causal conclusions from the results.

PISA dataset

Following Kelava et al. (2014), the first example uses the Jordan data of the Program for International Student Assessment 2006 dataset (Organisation for Economic Co-Operation and Development, 2009) for demonstration purposes. The sample consists of 6,038 15-year-old students, and the data measures their attitudes, motivation, and educational performance on a four-point scale. We use these data to run a simple model where enjoyment of science (ENJ, five items, x_1-x_5), academic self-concept in science (SC, six items, x_6-x_{11}), and their interaction (ENJ $\times$ SC) are used to explain career aspiration in science (four items, y_1-y_4). To simplify the analysis—given the mere didactical purpose of this example—we did not use survey weights and focused on a latent interaction effect rather than on estimating a full-polynomial model (like Kelava et al., 2014).

We estimate the LMS model in Mplus and both EXT-MATCHED-R and MIIV-PARCEL in R. Given the pedagogical nature of this example, we chose EXT-MATCHED over EXT-ALL, because it is considerably simpler to implement and can, thus, be a more useful blueprint for applied researchers interested using our code. We use $x_1x_6, x_2x_7, x_3x_8, x_4x_9$, and x_5x_{10} as product indicators for the latent interaction in the EXT-MATCHED-R model. We use x_1, x_6, and x_1x_6 to recast the latent interaction model into an observed variable one in MIIV-PARCEL, using x_2-x_5 to generate a parcel for ENJ, x_7-x_{11} to generate a parcel for SC, and the 20 possible interactions between the potential indicators of ENJ and SC (e.g., x_2x_7) to generate a parcel for ENJ × SC; we also use an index of career aspiration as a dependent variable for MIIV-PARCEL. We center all indicators before generating the product indicators.

Across estimators, we find positive estimates of both linear effects (see Table 2). However, the latent interaction effect estimates differ considerably between the estimators. According to LMS, the interaction term is negative and not significant (-.02, p = .344), whereas it is positive according to EXT (.05, p = .009) and MIIV (.08, p < .001); as a benchmark, the interaction term estimated with OLS (where scale means and their interaction are entered as predictors, ignoring measurement error) is positive, but not significant. To select which estimator to retain, we run the H_R specification test using EXT-MATCHED-R or MIIV-PARCEL as the consistent estimators. In both cases, we strongly reject the assumption of no distributional misspecification of LMS (EXT vs. LMS: H_R = 12.64, p < .001; EXT vs. MIIV: H_R = 20.43, p < .001), suggesting that the distributional assumptions of LMS are unlikely to be met. In this case, inspecting the distribution of the observed variables is not as useful because of the small number of response categories (four). However, both visual inspection and normality tests show that the observed variables are left skewed. Thus, this example shows that the conclusions drawn

using LMS may be wrong and that the specification tests can help researchers select the correct model.

[Table 2 about here]

ZVT dataset

For our second example, we use a smaller dataset where the independent variable is not a set of questionnaire responses but a psychometric test. Specifically, we use an unpublished dataset (available from the third author) based on surveys conducted in small and medium-sized firms in Kampala City (Uganda) from July 2012 to August 2012. In each structured questionnaire/interview, entrepreneurs were asked several questions, as well as to complete four versions of the Zahlen-Verbindungs-Test (ZVT), a trail-making test where subjects are asked to connect numbers ranging from 1 to 90 in four different matrixes with a hand-written line (Oswald & Roth, 1987). Crucially, the numbers are positioned randomly in each matrix and subjects have 30 seconds to complete each matrix; their score is given by the highest correct number they managed to connect within the time limit. The ZVT is, thus, a measure of information processing speed that correlates with measures of intelligence (e.g., Schweizer & Moosbrugger, 2004; Vernon, 1993).

We run a simple model where the network size of each entrepreneur (a measure of social capital) is explained by her/his intelligence and intelligence squared. We model this quadratic effect (i.e., an "interaction" of a variable with itself) because nonlinear effects of intelligence are becoming more and more acknowledged in the psychology literature (e.g., Antonakis et al., 2017; Gignac et al., 2020; Rammstedt et al., 2016). For MIIV-PARCEL, we use the first ZVT completed by the respondents as a reference indicator and all remaining indicators (or products between indicators) as parceled instruments. For EXT-MATCHED-R, we used all four ZVT versions to indicate the linear term and the squared term of each ZVT version as indicator for the

squared latent predictor. The results in Table 2 show that all estimators produce very similar results. According to all estimators, the linear (marginal) effect is negative and insignificant. Similarly, the quadratic term is negative and significant for LMS (-.01, p = .001) and for MIIV-PARCEL (-.01, p = .004), even though it is not significant—but similar in magnitude—for EXT-MATCHED (-.01, p = .057); as a benchmark, the OLS interaction estimate that neglects measurement error is negative and significant (-.01, p = .001).

Because all estimators return virtually the same result, choosing one or the other would lead to no substantive differences in conclusions. Calculating the H_R specification test using EXT-MATCHED or MIIV-PARCEL fails to reject the assumption of no misspecification of LMS in both cases. However, all multivariate normality tests for the observed indicators suggest that the distributions of the four ZVT versions are not multivariate normal.

In contrast to the PISA case, this example shows that the conclusions drawn by LMS may be correct and that our specification tests can help researchers select the appropriate estimator. Moreover, the example shows that relying on multivariate normality tests for the observed variable indicators might lead researchers to different conclusions. Of course, we cannot know whether LMS is biased in this case. However, the point estimates of LMS and both MIIV and EXT are virtually identical, suggesting that distributional misspecifications are not a major problem. It is also possible that the specification tests are underpowered because of the small sample size. Nevertheless, the example demonstrates that relying on multivariate normality tests alone would lead researchers to reject the significant LMS interaction estimate in favor of the non-significant EXT ones.

Discussion and Recommendations

Recommendations for practitioners

The general recommendation stemming from our results is simple: We strongly advise against putting faith in the LMS estimates unless the assumption that the latent predictors and the structural disturbances are normally distributed can be justified. Unfortunately, these assumptions cannot be tested directly. We show that the multivariate normality tests for the observed indicators of the predictor variables and the robust Hausman specification test are useful indirect tests of these assumptions. Based on our theoretical considerations and simulation results, we recommend using these two tools following the decision rule of Table 3.

When both the multivariate normality tests for the observed indicators and the specification test agree, the choice between LMS and an alternative estimator is clear. If neither the multivariate normality tests nor the specification test indicates a problem (i.e., normal indicators and no discernible difference between LMS and another consistent estimator), the researcher has some evidence that LMS estimates might be valid and can be used. In contrast, if both types of tests indicate a problem, researchers should discard the LMS results and trust consistent (but inefficient) alternatives, like EXT or MIIV. When the specification test and the multivariate normality tests disagree, the situation is more complicated. In such cases, we recommend trusting the specification test, especially if the sample size is reasonably large. This is because multivariate normality tests might fail to flag misspecifications caused by non-normal structural disturbances or indicate inconsequential distributional misspecifications due to non-normal measurement errors.

The robust Hausman test is no panacea, however. In small samples, the specification test is underpowered. To address this issue, and given the expected losses related to Type II error (i.e., using an inconsistent estimator) and Type I error (i.e., using an inefficient estimator), we recommend an $\alpha = 10\%$ significance threshold for this test (cf. Kim & Choi, 2021; Maier & Lakens, 2022). With this more lenient criterion, LMS models can be correctly flagged as

misspecified much more often, including in conditions with mild skewness. Even with this more lenient significance level, the Type I error rate remains extremely low: The robust Hausman tests flag as misspecified a correct LMS model—at worse—with 3% probability (see Supplementary Material 1, 2, 3, and 4 for the results with $\alpha = 10\%$). On the contrary, the "lower bound" Hausman test is too underpowered to be practically useful unless the sample size is in the thousands.

Finally, readers must be mindful that although multivariate normality tests and the specification test can uncover potential distributional misspecifications, they are not tests for structural misspecifications (e.g., omitted paths, unobserved confounders, incorrect measurement models). Thus, the tests can only complement clear theory, a defensible causal identification strategy, and a careful examination of the fit of the measurement model.

[Table 3 about here]

Recommendations for future research

Our study suggests at least two areas where further research is needed. First, our simulation on non-normal structural disturbances shows patterns that prior research has—to the best of our knowledge—never documented. This simulation revealed a particularly bizarre behavior of the LMS estimator, highlighting a bimodal sampling distribution and a catastrophic bias even when the sample size was in the thousands. Although we heavily diagnosed the problem and speculated about its origin, future technical work should focus on understanding why this pattern of results emerges. Second, we developed the specification tests for continuous variable estimators and the single parameter case, which is simple to calculate. Future research could focus on developing multiparameter tests or tests that support comparing the results from a continuous-variables estimator against categorical-variable estimation results.

Finally, this work is not without limitations. As with all simulations, it is impossible to address every scenario that might be relevant for applied researchers. For instance, we did not examine models with multiple dependent variables. We also did not consider all possible non-normality shapes or models with many indicators or unequal factor loading. Similarly, we did not review and study all possible multivariate normality tests and latent interaction estimators. Most notably, we did not study the "nonlinear structural equation mixture model" approach (NSEMM, Kelava et al., 2014). This estimator is promising yet relatively unknown in applied research. Although it should be more robust to non-normality than LMS and be as efficient under normality, it is also computationally costly and can be complex to implement (see the companion code in Kelava & Brandt, 2022). NSEMM also comes with its own set of assumptions and limitations (i.e., the properties of NSEMM are ensured if the correct number of latent classes is used, and simulation evidence suggests that its robustness properties are valid only under relatively strong reliability, see Brandt et al., 2020). Thus, a particularly interesting question is whether the specification test we propose can be used relying on NSEMM (rather than EXT or MIIV) as a consistent estimator or if the specification test could be applied to detect potential bias due to latent class misspecifications in NSEMM. Future simulation work will, hopefully, fill these and other gaps[13].

Conclusion

Normality plays a key role when estimating latent interaction effects with the commonly used LMS approach (i.e., `XWITH` in Mplus). However, the normality assumption is rarely

[13] We run NSEMM with two mixture components for the practical examples we discuss. In the PISA dataset, NSEMM produces results that are qualitatively similar to LMS (interaction effect: .02, SE = .02, p = .473). For the ZVT dataset, NSEMM produces visibly different results compared to both LMS and the other approaches (interaction effect: -.002, SE = .001, p = .054).

discussed by practitioners, let alone tested for. Our theoretical considerations and simulation results provide clear methodological guidelines that can help applied researchers to understand if, when, and how violations of normality can be a major cause of concern for their latent interaction models and how possible problems can be detected with appropriate statistical tests. We recommend that researchers always test latent variable normality using both multivariate normality tests of the observed indicators of the latent predictors and the specification tests we proposed in this paper. Failure to do so may mean publishing potentially misleading estimates.

References

Aguinis, H., & O'Boyle Jr, E. (2014). Star performers in twenty-first century organizations. *Personnel Psychology*, *67*(2), 313-350.

Aguirre-Urreta, M., Rönkkö, M., & Hu, J. (2020). Polynomial Regression in IS Research: Issues and Consequences. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, *51*(3), 55-80.

Aït-Sahalia, Y., & Xiu, D. (2019). A Hausman test for the presence of market microstructure noise in high frequency data. *Journal of Econometrics*, *211*(1), 176-205.

Alessandri, G., Cortina, J. M., Sheng, Z., & Borgogni, L. (2021). Where you came from and where you are going: The role of performance trajectory in promotion decisions. *Journal of Applied Psychology*, *106*(4), 599-623.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, *21*(6), 1086-1120.

Antonakis, J., House, R. J., & Simonton, D. K. (2017). Can super smart leaders suffer from too much of a good thing? The curvilinear effect of intelligence on perceived leadership behavior. *Journal of Applied Psychology*, *102*(7), 1003-1021.

Aytürk, E., Cham, H., Jennings, P. A., & Brown, J. L. (2020). Latent Variable Interactions With Ordered-Categorical Indicators: Comparisons of Unconstrained Product Indicator and Latent Moderated Structural Equations Approaches. *Educational and Psychological Measurement*, *80*, 262–292.

Aytürk, E., Cham, H., Jennings, P. A., & Brown, J. L. (2021). Exploring the performance of latent moderated structural equations approach for ordered-categorical items. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(3), 410-422.

Baum, C. F., Schaffer, M. E., & Stillman, S. (2003). Instrumental variables and GMM: Estimation and testing. *The Stata Journal*, *3*(1), 1-31.

Belogolovsky, E., Bamberger, P., Alterman, V., & Wagner, D. T. (2016). Looking for assistance in the dark: Pay secrecy, expertise perceptions, and efficacious help seeking among members of newly formed virtual work groups. *Journal of Business and Psychology*, *31*, 459-477.

Blake, K. R., & Gangestad, S. (2020). On attenuated interactions, measurement error, and statistical power: Guidelines for social and personality psychologists. *Personality and Social Psychology Bulletin*, *46*(12), 1702-1711.

Bollen, K. A. (1996). An alternative two stage least squares (2SLS) estimator for latent variable equations. *Psychometrika*, *61*(1), 109-121.

Bollen, K. A., Fisher, Z. F., Giordano, M. L., Lilly, A. G., Luo, L., & Ye, A. (2022). An introduction to model implied instrumental variables using two stage least squares (MIIV-2SLS) in structural equation models (SEMs). *Psychological Methods*, *27*(5), 752–772.

Bollen, K. A., Kirby, J. B., Curran, P. J., Paxton, P. M., & Chen, F. (2007). Latent variable models under misspecification: Two-stage least squares (2SLS) and maximum likelihood (ML) estimators. *Sociological Methods & Research*, *36*(1), 48-86.

Bollen, K. A., Kolenikov, S., & Bauldry, S. (2014). Model-implied instrumental variable—generalized method of moments (MIIV-GMM) estimators for latent variable models. *Psychometrika*, *79*(1), 20-50.

Bollen, K. A., & Paxton, P. (1998). Interactions of latent variables in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(3), 267-293.

Boomsma, A. (2013). Reporting Monte Carlo studies in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(3), 518-540.

Brandt, H., Kelava, A., & Klein, A. (2014). A simulation study comparing recent approaches for the estimation of nonlinear effects in SEM under the condition of nonnormality. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(2), 181-195.

Brandt, H., Umbach, N., Kelava, A., & Bollen, K. A. (2020). Comparing estimators for latent interaction models under structural and distributional misspecifications. *Psychological Methods*, *25*(3), 321–345.

Browne, M. W., & Shapiro, A. (1988). Robustness of normal theory methods in the analysis of linear latent variate models. *British Journal of Mathematical and Statistical Psychology*, *41*(2), 193-208.

Cain, M. K., Zhang, Z., & Yuan, K.-H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*, *49*(5), 1716-1735.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.

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*(6), 840-876.

Cheung, G. W., & Lau, R. S. (2017). Accuracy of parameter estimates and confidence intervals in moderated mediation models: A comparison of regression and latent moderated structural equations. *Organizational Research Methods*, *20*(4), 746-769.

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, *19*(2), 300-315.

Cortina, J. M., Markell-Goldstein, H. M., Green, J. P., & Chang, Y. (2021). How Are We Testing Interactions in Latent Variable Models? Surging Forward or Fighting Shy? *Organizational Research Methods*, *24*(1), 26-54.

Creel, M. (2004). Modified Hausman tests for inefficient estimators. *Applied Economics*, *36*(21), 2373-2376.

Culpepper, S. A. (2012). Evaluating EIV, OLS, and SEM estimators of group slope differences in the presence of measurement error: The single-indicator case. *Applied Psychological Measurement*, *36*(5), 349-374.

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*(1), 16-29.

Dawson, J. F. (2014). Moderation in management research: What, why, when, and how. *Journal of Business and Psychology*, *29*(1), 1-19.

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

Doornik, J. A., & Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics*, *70*, 927-939.

Fair, R. C., & Parke, W. R. (1980). Full-information estimates of a nonlinear macroeconometric model. *Journal of Econometrics*, *13*(3), 269-291.

Fasbender, U., & Gerpott, F. H. (2023). Designing work for change and its unintended side effects. *Journal of Vocational behavior*, *145*, 103913.

Fischer, T., Tian, A. W., Lee, A., & Hughes, D. J. (2021). Abusive supervision: a systematic review and fundamental rethink. *The Leadership Quarterly*, *32*(6), 101540.

Foldnes, N., & Grønneberg, S. (2015). How general is the Vale–Maurelli simulation approach? *Psychometrika*, *80*(4), 1066-1083.

Foldnes, N., & Hagtvet, K. A. (2014). The choice of product indicators in latent variable interaction models: Post hoc analyses. *Psychological Methods*, *19*(3), 444-457.

Gignac, G. E., Walker, B., Burtenshaw, T., & Fay, N. (2020). On the nonlinear association between intelligence and openness: Not much of an effect beyond an average IQ. *Personality and Individual Differences*, *166*, 110169.

Hahn, J., & Hausman, J. (2002). A new specification test for the validity of instrumental variables. *Econometrica*, *70*(1), 163-189.

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in M plus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621-638.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, *46*(6), 1251-1271.

Henze, N., & Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in statistics-Theory and Methods*, *19*(10), 3595-3617.

Katz, J. N., & Katz, G. (2010). Correcting for survey misreports using auxiliary information with an application to estimating turnout. *American Journal of Political Science*, *54*(3), 815-835.

Kelava, A., & Brandt, H. (2009). Estimation of nonlinear latent structural equation models using the extended unconstrained approach. *Review of Psychology*, *16*(2), 123-132.

Kelava, A., & Brandt, H. (2022). Latent Interaction Effects. *Handbook of Structural Equation Modeling*, 427-446.

Kelava, A., & Nagengast, B. (2012). A Bayesian model for the estimation of latent interaction and quadratic effects when latent variables are non-normally distributed. *Multivariate Behavioral Research*, *47*(5), 717-742.

Kelava, A., Nagengast, B., & Brandt, H. (2014). A nonlinear structural equation mixture modeling approach for nonnormally distributed latent predictor variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(3), 468-481.

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, *96*(1), 201-210.

Kim, J. H., & Choi, I. (2021). Choosing the Level of Significance: A Decision-theoretic Approach. *Abacus*, *57*(1), 27-71.

Kleiber, C., & Zeileis, A. (2009). *AER: Applied Econometrics with R. R package version 1.1*.

Klein, A. G., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, *65*(4), 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*(4), 647-673.

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling*. Guilford publications.

Korkmaz, S., Göksülük, D., & Zararsiz, G. (2014). MVN: An R package for assessing multivariate normality. *R Journal*, *6*(2), 151-162.

Lodder, P., Denollet, J., Emons, W. H., Nefs, G., Pouwer, F., Speight, J., & Wicherts, J. M. (2019). Modeling interactions between latent variables in research on Type D personality: A Monte Carlo simulation and clinical study of depression and anxiety. *Multivariate Behavioral Research*, *54*(5), 637-665.

Maier, M., & Lakens, D. (2022). Justify your alpha: A primer on two practical approaches. *Advances in Methods and Practices in Psychological Science*, *5*(2), 25152459221080396.

Mair, P., Satorra, A., & Bentler, P. M. (2012). Generating nonnormal multivariate data using copulas: Applications to SEM. *Multivariate Behavioral Research*, *47*(4), 547-565.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*(3), 519-530.

Marsh, H. W., Wen, Z., & Hau, K.-T. (2004). Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction. *Psychological Methods*, *9*(3), 275-300.

Marsh, H. W., Wen, Z., Hau, K.-T., & Nagengast, B. (2006). Structural equation models of latent interaction and quadratic effects. *Structural equation modeling: A second course*, 225-265.

McDermott, R. C., Berry, A. T., Borgogna, N. C., Cheng, H.-L., Wong, Y. J., Browning, B., & Carr, N. (2020). Revisiting the paradox of hope: The role of discrimination among first-year Black college students. *Journal of Counseling Psychology*, *67*(5), 637.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156-166.

Millimet, D. L., & Parmeter, C. F. (2022). Accounting for skewed or one-sided measurement error in the dependent variable. *Political Analysis*, *30*(1), 66-88.

Mooijaart, A., & Bentler, P. M. (2010). An alternative approach for nonlinear latent variable models. *Structural Equation Modeling*, *17*(3), 357-373.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074-2102.

Moulder, B. C., & Algina, J. (2002). Comparison of methods for estimating and testing latent variable interactions. *Structural Equation Modeling*, *9*(1), 1-19.

Olvera Astivia, O. L., & Zumbo, B. D. (2015). A cautionary note on the use of the Vale and Maurelli method to generate multivariate, nonnormal data for simulation purposes. *Educational and Psychological Measurement*, *75*(4), 541-567.

Organisation for Economic Co-Operation and Development. (2009). *PISA 2006: Science Competencies for Tomorrow's World* (Technical report, OECD, Issue.

Oswald, W. D., & Roth, E. (1987). *Der Zahlen-Verbindungs-Test (ZVT)*. Hogrefe Verlag fuer Psychologie.

Rammstedt, B., Danner, D., & Martin, S. (2016). The association between personality and cognitive ability: Going beyond simple effects. *Journal of Research in Personality*, *62*, 39-44.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354-373.

Rönkkö, M., McIntosh, C., & Antonakis, J. (2020). On the Small-Sample Properties of "Consistent Partial Least Squares for Nonlinear Structural Equation Models.". *Working paper*.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software*, *48*(2), 1-36.

Royston, P. (1992). Approximating the Shapiro-Wilk W-test for non-normality. *Statistics and computing*, *2*, 117-119.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507-514.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, *1*(2), 199-223.

Schwarz, N. (1999). Self-reports: how the questions shape the answers. *American Psychologist*, *54*(2), 93-105.

Schweizer, K., & Moosbrugger, H. (2004). Attention and working memory as predictors of intelligence. *Intelligence*, *32*(4), 329-347.

Su, R., Zhang, Q., Liu, Y., & Tay, L. (2019). Modeling congruence in organizational research with latent moderated structural equations. *Journal of Applied Psychology*, *104*(11), 1404–1433.

Sun, X., Hall, G. C. N., DeGarmo, D. S., Chain, J., & Fong, M. C. (2021). A longitudinal investigation of discrimination and mental health in Chinese international students: The role of social connectedness. *Journal of Cross-Cultural Psychology*, *52*(1), 61-77.

Taggart, T. C., Bannon, S. M., & Hammett, J. F. (2019). Personality traits moderate the association between conflict resolution and subsequent relationship satisfaction in dating couples. *Personality and Individual Differences*, *139*, 281-289.

Umbach, N., Naumann, K., Brandt, H., & Kelava, A. (2017). Fitting nonlinear structural equation models in R with package nlsem. *Journal of Statistical Software*, *77*(i07).

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, *48*(3), 465-471.

Vernon, P. A. (1993). Der Zahlen-Verbindungs-Test and other trail-making correlates of general intelligence. *Personality and Individual Differences*, *14*(1), 35-40.

Wall, M. M., & Amemiya, Y. (2001). Generalized appended product indicator procedure for nonlinear structural equation analysis. *Journal of Educational and Behavioral Statistics*, *26*(1), 1-29.

Wall, M. M., & Amemiya, Y. (2003). A method of moments technique for fitting interaction effects in structural equation models. *British Journal of Mathematical and Statistical Psychology*, *56*(1), 47-63.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817-838.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*(1), 1-25.

Wooldridge, J. M. (2002). *Introductory econometrics: A modern approach*. South-Western, Div of Thomson Le.

Yuan, K.-H., Bentler, P. M., & Zhang, W. (2005). The effect of skewness and kurtosis on mean and covariance structure analysis: The univariate case and its multivariate implication. *Sociological Methods & Research*, *34*(2), 240-258.

Zeckhauser, R., & Thompson, M. (1970). Linear regression with non-normal error terms. *The Review of Economics and Statistics*, 280-286.

Tables

Table 1

Overview of Monte Carlo simulations and main results, broken down by most significant experimental condition where relevant.

Simulation	Non-normality in	Effects on LMS bias	Reject biased LMS (Power)		Reject unbiased LMS (False positives)		Explanation
			MVN tests	H_R test	MVN tests	H_R test	
#1	Latent variables *(interaction effect in the population)*	Very large	82%	41% [49%]	9%	0% [0%]	MVN tests are more powerful than the specification test
	Latent variables *(no interaction effect)*	Large	83%	30% [42%]	56%	0% [1%]	MVN tests can flag deviations from normality that are inconsequential for LMS bias
#2	Continuous measurement errors	None	NA[a]	NA[a]	89%	0% [0%]	The specification test rarely signals a misspecification in LMS MVN flags as misspecified most LMS models where observed indicators are non-normally distributed
#3	Latent variables & categorical indicators	Extremely large	71%	42% [46%]	65%	0% [0%]	MVN tests are more powerful than the specification test Some MVN tests can flag deviations from normality that are inconsequential for LMS bias
#4	Structural disturbance *(high reliability)*	Little-to-no bias	5%	6% [9%]	5%	0% [1%]	MVN test cannot detect these misspecifications
	Structural disturbance *(low reliability)*	Potentially catastrophic	5%	40% [42%]	5%	1% [2%]	MVN tests can never detect non-normality in the structural disturbance, different from the specification test

Note. Share of LMS models rejected on average by the three multivariate normality (MVN) tests and H_R using EXT-ALL-R as consistent estimator. Average of all conditions with $N >= 350$. An LMS model is considered "biased" if the absolute bias of LMS > |.05|; it is considered "unbiased" otherwise. α = 5%, [α = 10% in brackets].

[a] = no LMS model has an absolute bias larger than |.05| in this condition.

Table 2

Empirical example based on the PISA dataset (Panel A) and ZVT dataset (Panel B).

Panel A – Dependent variable: career aspiration in science				
	LMS	MIIV-PARCEL	EXT-MATCHED-R	OLS
ENJ	.517***	.476***	. 528***	.429***
	(.019)	(.021)	(.022)	(.012)
SC	.469***	.492***	.475***	.369***
	(.023)	(.027)	(.025)	(.013)
$ENJ \times SC$	-.019	.075***	.053**	.019
	(.020)	(.021)	(.020)	(.012)
H_R (LMS vs. EXT-MATCHED-R)	12.644***			
H_R (LMS vs. MIIV-PARCEL)	20.429***			
Mardia skewness test	6407.428***			
Mardia kurtosis test	122.625***			
Doornik-Hansen test	3310.517***			
Panel B – Dependent variable: network size				
	LMS	MIIV-PARCEL	EXT-MATCHED-R	OLS
ZVT	-.028	-.027	-.028	-.020
	(.022)	(.023)	(.021)	(.017)
ZVT^2	-.008**	-.008**	-.007	-.005**
	(.002)	(.003)	(.004)	(.002)
H_R (LMS vs. EXT-MATCHED-R)	.027			
H_R (LMS vs. MIIV-PARCEL)	.010			
Mardia skewness test	482.391***			
Mardia kurtosis test	31.916***			
Doornik-Hansen test	101.547***			

Note. *** $p < .001$, ** $p < .01$, * $p < .05$. N = 6,038 for Panel A; N = 219 for Panel B. Constant terms omitted from the table. Standard errors in parentheses. LMS = Latent Moderated Structural Equations; EXT-MATCHED-R = Extended Unconstrained approach using 3 matched indicators with robust standard errors; MIIV-PARCEL = Model-Implied Instrumental variables estimator using parcels of all available instruments; OLS = Ordinary Least Squares using all indicators as indexes; H_R (LMS vs. EXT-MATCHED-R) = robust Hausman specification test contrasting LMS and EXT-MATCHED-R; H_R (LMS vs. MIIV-PARCEL) = robust Hausman specification test contrasting LMS and MIIV-PARCEL; Mardia skewness test, Mardia kurtosis test, and Doornik-Hansen test are multivariate normality tests for the observed indicators of the latent predictor(s).

Table 3

Using multivariate normality and specification tests' results for model selection.

		Result of the robust Hausman test comparing LMS against EXT or MIIV	
		Non-significant ($p > .10$)	Significant ($p < .10$)
Multivariate normality of observed indicators of the predictor latent variables	Supported	Lack of multivariate normality is not a problem and the LMS results should be reported.	This may indicate a non-normal disturbance term, which can severely bias the LMS results. LMS should *not* be used.
	Not supported	If there is a clear reason for why the normality tests fail (e.g., categorical indicators) and the sample size is sufficiently large (e.g., 350 or more), LMS can be used.	Lack of multivariate normality is clearly a problem and LMS should *not* be used.

Figures

Figure 1

Graphical representation of a simple latent interaction model.

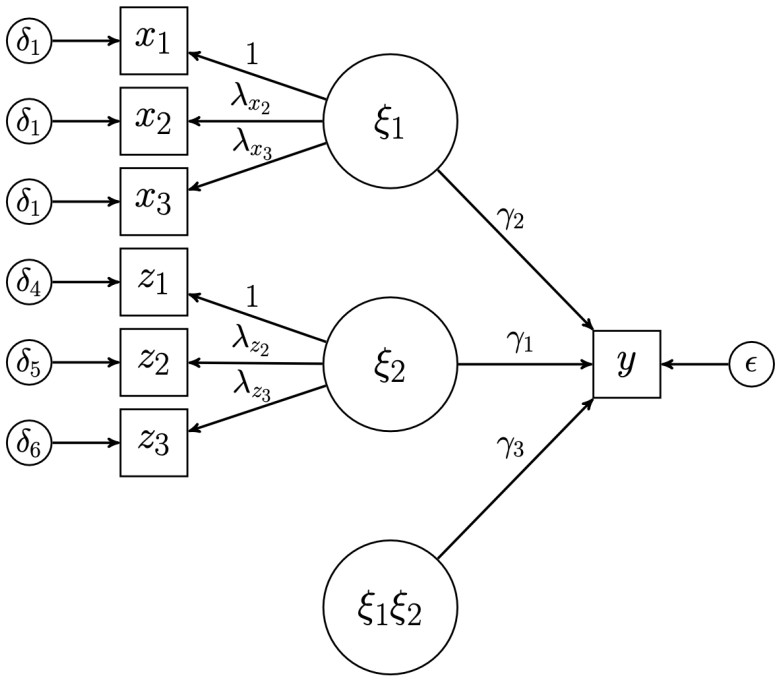
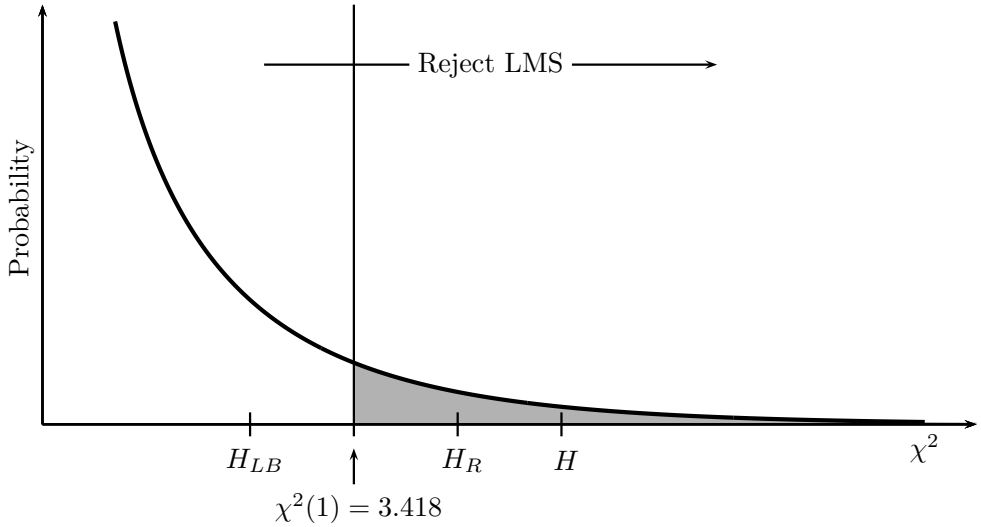


Figure 2

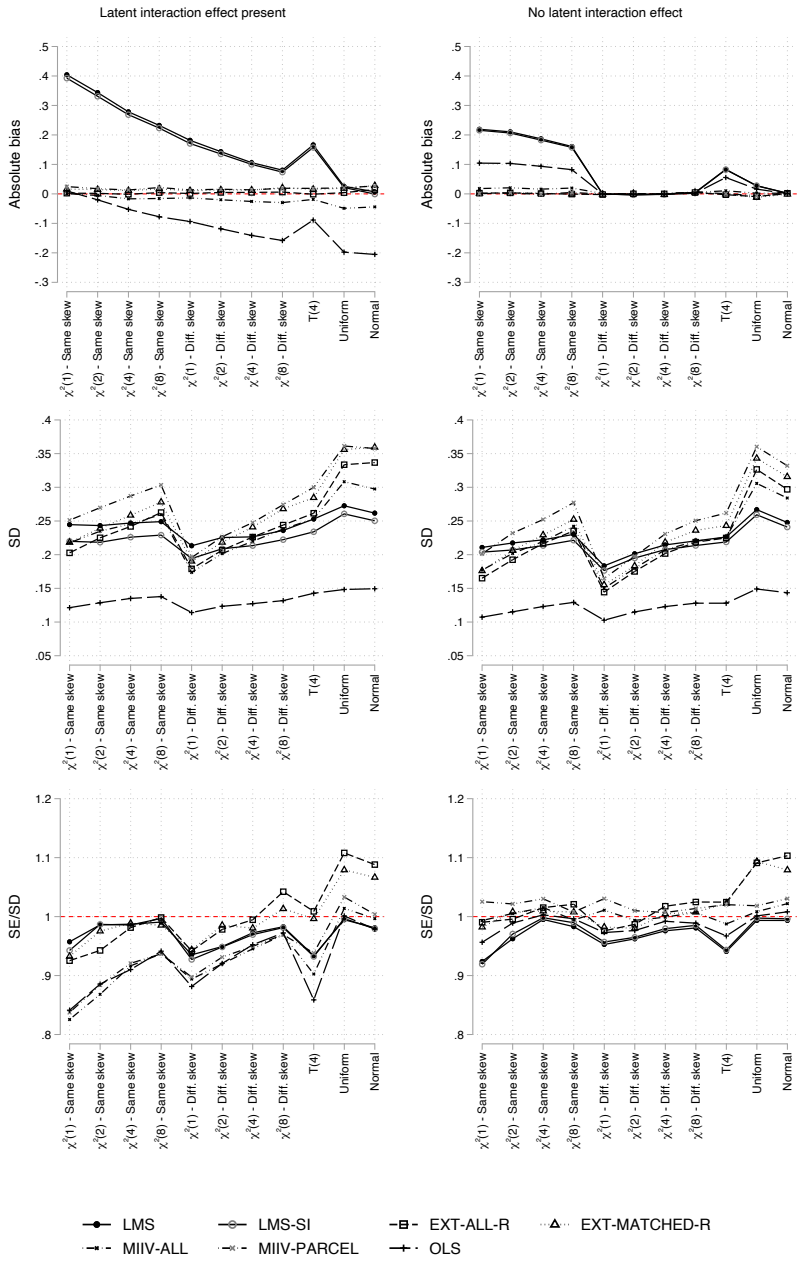
Specification tests and rejection regions under the null hypothesis of no distributional misspecification of LMS.



Note. When the null hypothesis is correct, $H_{LB} < H_R < H$. Thus, for the 95% critical value of $\chi^2(1) = 3.418$, the probability of rejecting LMS is highest with the statistic H and lowest with the statistic H_{LB}.

Figure 3

Simulation 1: Absolute bias, standard deviation, and standard error over standard deviation over latent predictors' distribution and presence/absence of the interaction effect.

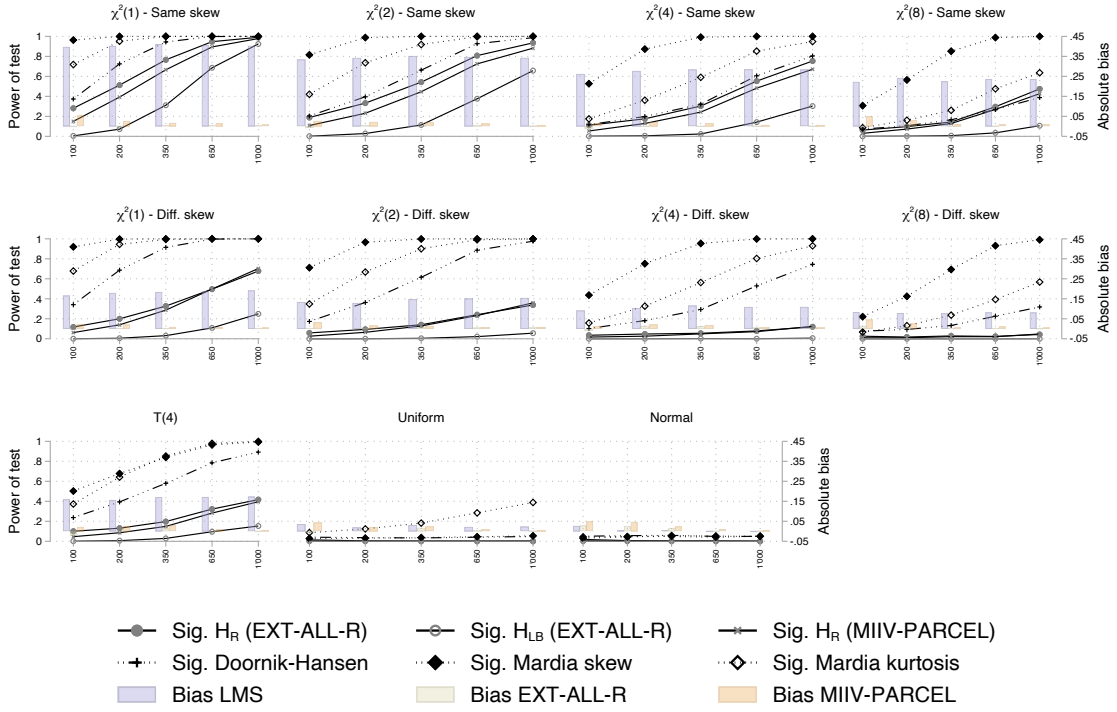


Note. LMS = Latent Moderated Structural Equations; LMS-SI = single-indicator Latent Moderated Structural Equations; EXT-MATCHED-R = Extended Unconstrained approach using 3 matched indicators with robust standard errors; EXT-ALL-R = Extended Unconstrained approach using all potential indicators with robust standard errors; MIIV-ALL = Model-Implied Instrumental variables estimator using all available instruments; MIIV-PARCEL = Model- Implied Instrumental

variables estimator using parcels of all available instruments; OLS = Observed-variable regression model via Ordinary Least Squares using scale means and their interactions as predictors. Absolute bias = $\frac{1}{R}\sum_{i=1}^{R}\hat{\gamma}_{3i} - \gamma_3$; SD = $\frac{1}{R-1}\sum_{i=1}^{R}\left(\hat{\gamma}_{3i} - \bar{\hat{\gamma}}_3\right)$; SE = $\sqrt{\frac{1}{R}\sum_{i=1}^{R}\widehat{Var}(\hat{\gamma}_{3i})}$, where R is the number of successful replications in each condition.

Figure 4

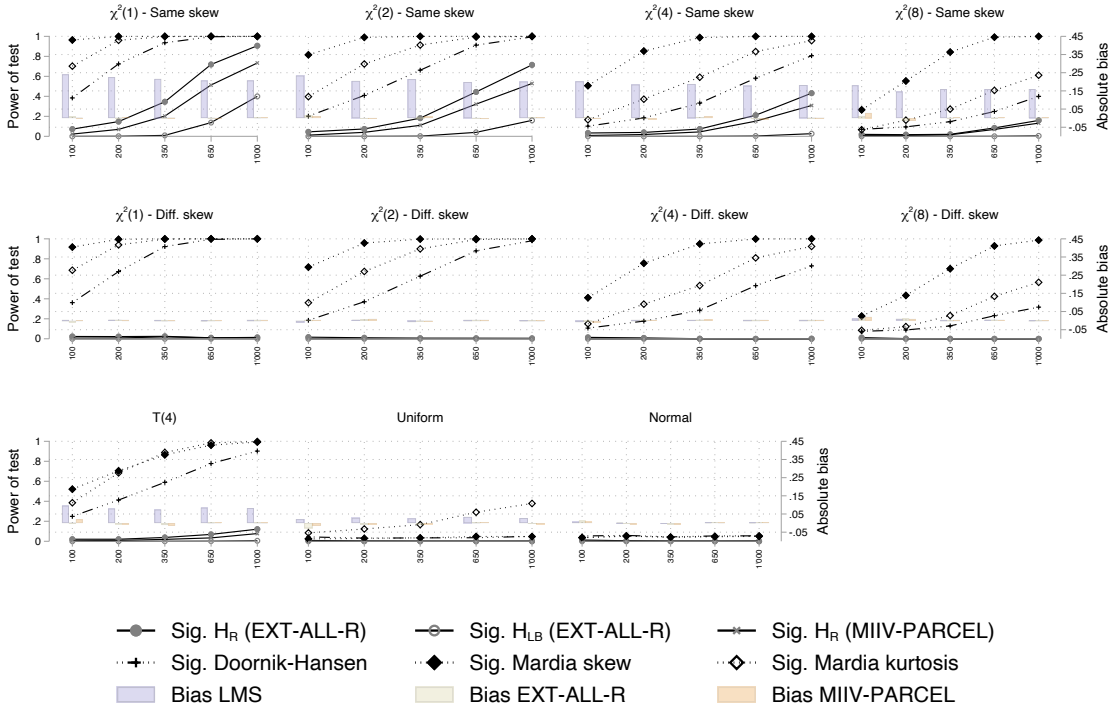
Simulation 1: Absolute bias of estimators and statistical power of specification tests over sample size by latent predictors' distribution when the interaction effect is set to .50.



Note. Statistical power at $\alpha = 5\%$. LMS = Latent Moderated Structural Equations; EXT-ALL-R = Extended Unconstrained approach using all potential indicators with robust standard errors; MIIV-PARCEL = Model-Implied Instrumental variables estimator using parcels of all available instruments.

Figure 5

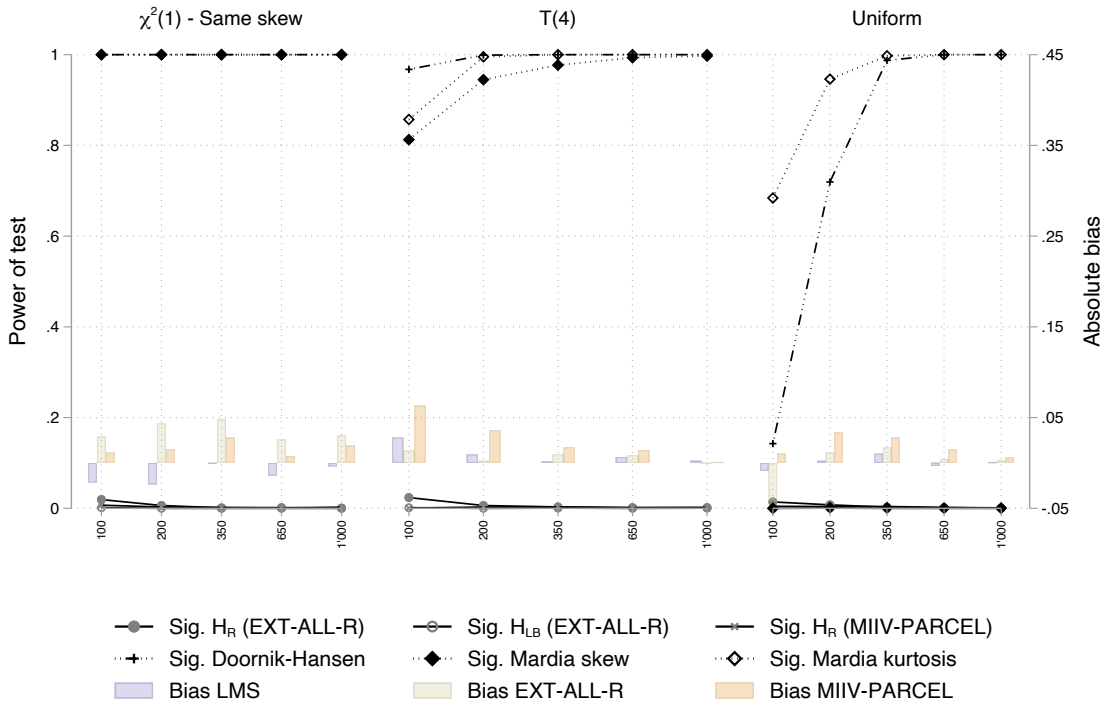
Simulation 1: Absolute bias of estimators and statistical power of specification tests over sample size by latent predictors' distribution when the interaction effect is set to 0.



Note. Statistical power at $\alpha = 5\%$. LMS = Latent Moderated Structural Equations; EXT-ALL-R = Extended Unconstrained approach using all potential indicators with robust standard errors; MIIV-PARCEL = Model-Implied Instrumental variables estimator using parcels of all available instruments.

Figure 6

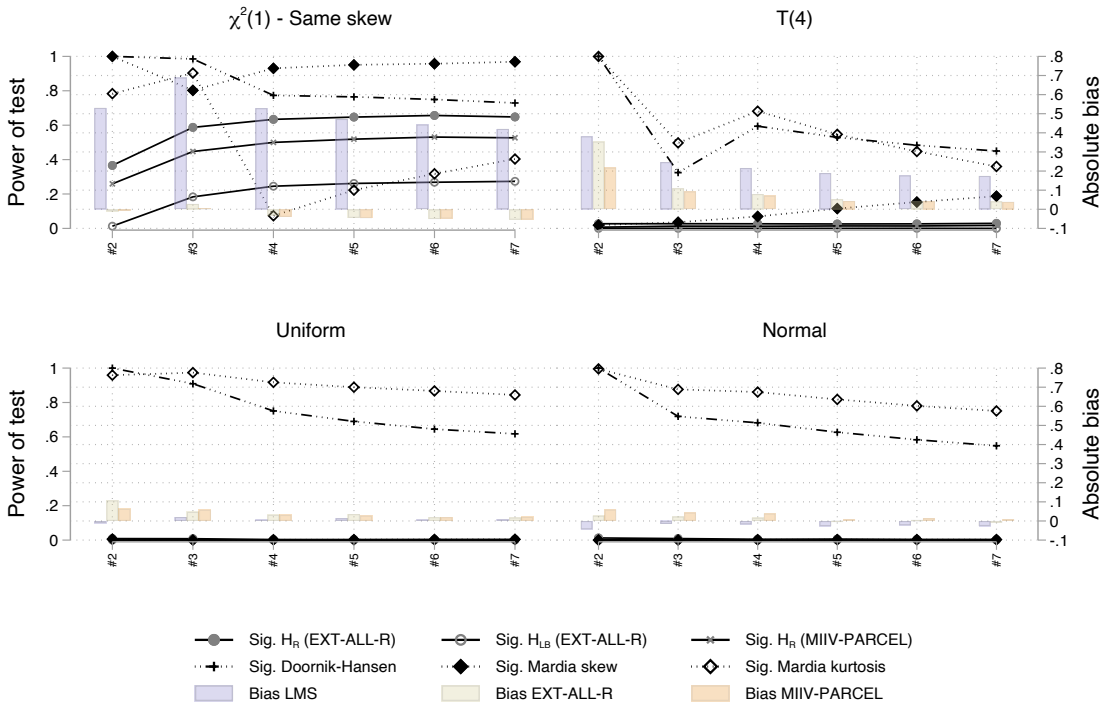
Simulation 2: Absolute bias of estimators and statistical power of specification tests with normal latent variables over measurement errors' distributions.



Note. Statistical power at $\alpha = 5\%$. LMS = Latent Moderated Structural Equations; EXT-ALL-R = Extended Unconstrained approach using all potential indicators with robust standard errors; MIIV-PARCEL = Model-Implied Instrumental variables estimator using parcels of all available instruments. In all conditions, latent predictors are normally distributed.

Figure 7

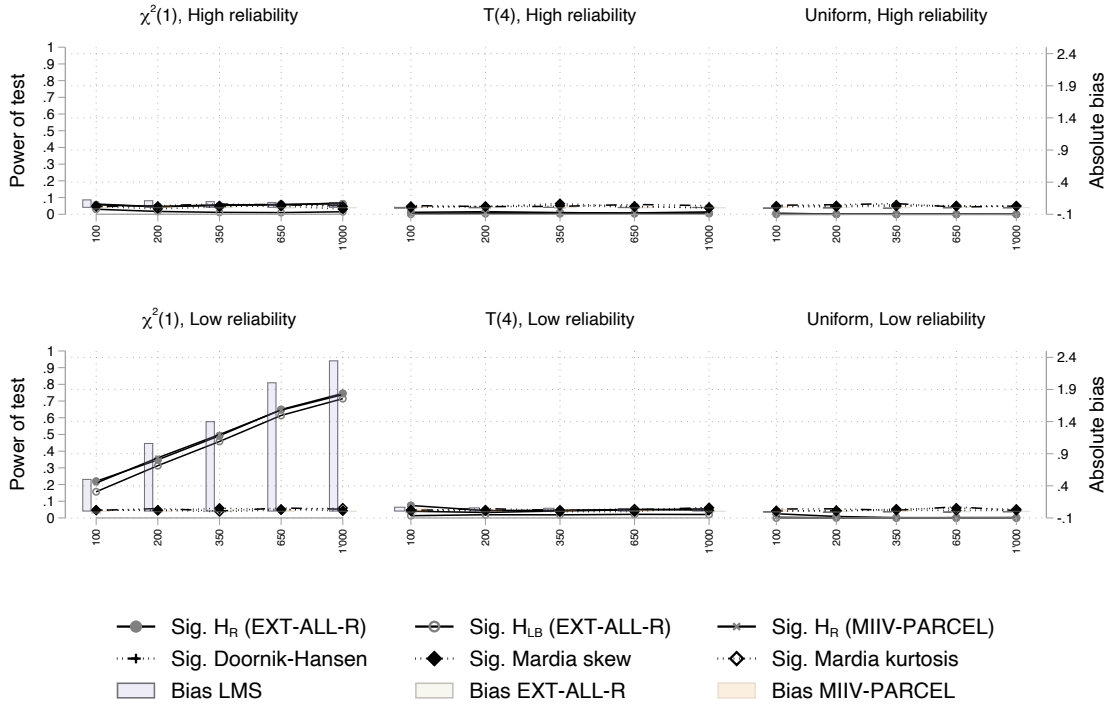
Simulation 3: Absolute bias of estimators and statistical power of specification tests over the number of observed indicators' categories by latent predictors' distribution.



Note. Statistical power at $\alpha = 5\%$. LMS = Latent Moderated Structural Equations; EXT-ALL-R = Extended Unconstrained approach using all potential indicators with robust standard errors; MIIV-PARCEL = Model-Implied Instrumental variables estimator using parcels of all available instruments.

Figure 8

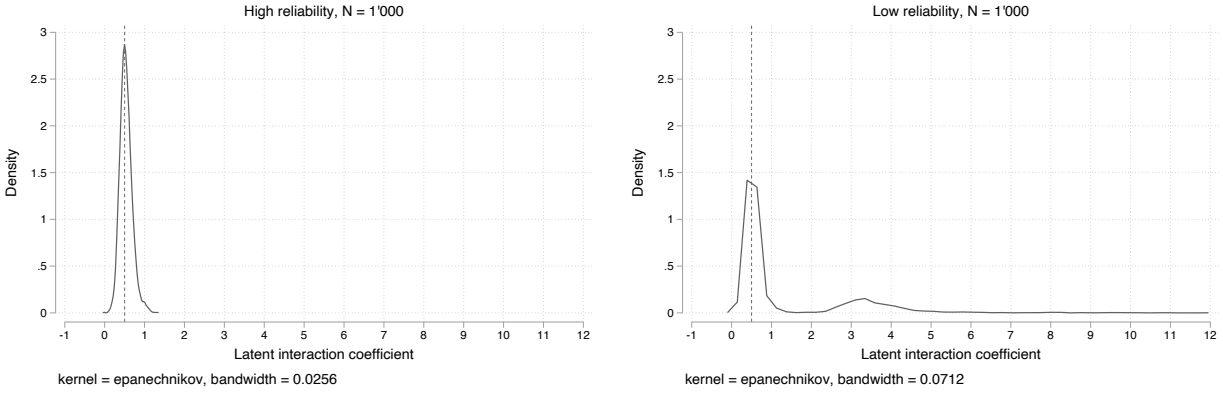
Simulation 4: Absolute bias of estimators and statistical power of specification tests with normal latent variables over disturbance's distributions.



Note. Statistical power at $\alpha = 5\%$. LMS = Latent Moderated Structural Equations; EXT-ALL-R = Extended Unconstrained approach using all potential indicators with robust standard errors; MIIV-PARCEL = Model-Implied Instrumental variables estimator using parcels of all available instruments. In all conditions, latent predictors are normally distributed.

Figure 9

Simulation 4: Distribution of the estimated latent interaction effect with LMS in the largest sample size condition when the structural disturbance is $\chi^2(1)$ *distributed, by reliability of the indicators.*



Note. High reliability = standard deviation of indicators' measurement errors equals .86; Low reliability = standard deviation of indicators' measurement errors equals 1.14.