SIX METHODS FOR LATENT MODERATION ANALYSIS IN MARKETING RESEARCH: A COMPARISON AND GUIDELINES

Pre-print November 2, 2021

~~Under peer review, please do not share or cite~~

Forthcoming, Journal of Marketing Research

Constant Pieters[a,*]

Rik Pieters[a]

Aurélie Lemmens[b]

[a] Department of Marketing, Tilburg School of Economics and Management, Tilburg University, Warandelaan 2, PO Box 90153, 5000 LE Tilburg, The Netherlands

[b] Department of Marketing Management, Rotterdam School of Management, Erasmus University Rotterdam, PO Box 1738, 3000 DR Rotterdam, The Netherlands

[*] Corresponding author.
E-mail addresses: c.pieters@tilburguniversity.edu (C. Pieters), f.g.m.pieters@tilburguniversity.edu (R. Pieters), lemmens@rsm.nl (A. Lemmens)

SIX METHODS FOR LATENT MODERATION ANALYSIS IN MARKETING RESEARCH: A COMPARISON AND GUIDELINES

ABSTRACT

Moderation analysis investigates the conditions under which variables affect outcomes. In marketing, it is common that at least one of the target moderation variables is latent and measured by multiple indicators with measurement error. This paper compares six methods for latent moderation analysis: multi-group, means, corrected means, factor scores, product indicators, and latent product. It reviews their use in marketing research, describes their assumptions, and compares their performance with Monte Carlo simulations. Several recommendations follow from the results. First, although the means method is the most frequently used method in the review (94% of articles), it should only be used when reliabilities of the moderation variables are close to one, which is rare. In that situation, all methods except the multi-group method perform similarly well. Second, the results support the use of the factor scores method and latent product method when reliabilities are smaller than one. These methods perform best with parameter and standard error bias $\leq 5\%$ under most investigated conditions. Third, specific settings can warrant the use of the multi-group method (if the moderator is discrete), the corrected means method (if moderation variables are single-indicators) and the product indicators method (if indicators are non-normally distributed). Practical guidelines and sample code on four statistical platforms are provided to stimulate the adoption of best practices for latent moderation analysis.

Keywords: moderation analysis, measurement error, research methods

Investigating the boundary conditions to a phenomenon is central to academic research and crucial for decision-makers. In marketing, it commonly involves a latent moderation analysis in which at least one of the target moderation variables is latent and is measured by one or more reflective indicators. Recent examples include Auh et al. (2019), who showed that customer orientation dampens the effect of customer participation on satisfaction (all three variables are latent). Another example is a study by Atasoy and Morewedge (2017) that found greater differences in (latent) perceptions of psychological ownership between physical and digital books (manipulated) when consumers had a stronger need for control (latent trait).

This paper focuses on latent moderation analysis and compares six main methods that differ in their approach and assumptions: the multi-group, means, corrected means, factor scores, product indicators, and latent product methods. Table 1 summarizes a literature review of 1,144 articles published in the *Journal of Marketing Research, Journal of Marketing, Journal of Consumer Research* and *Marketing Science* between 2015 and 2019. It shows that methods have not been equally popular. Among 656 estimated moderation effects in 164 articles, 94% of articles used the means method.

The means method takes unit weighted mean scores of the indicators without accounting for the remaining measurement error in the score. Measurement error is the difference between observed and true values of a score (Wooldridge 2015, p. 288). Its magnitude is determined by one minus the score's reliability, which is the proportion of systematic variance in the score with respect to its total variance (Bollen 1989, p. 156). It is known that not accounting for measurement error can severely bias estimates and/or standard errors (Bollen 1989; Cohen et al. 2003; Grewal et al. 2004; Spearman 1904; Wooldridge 2015). Bias is the difference between estimated and true values of a parameter or its standard error (Wooldridge 2015). Thus, the popularity of the means method is in stark contrast with its reported poor statistical properties in face of measurement error.

Nevertheless, multiple reasons can explain the common use of the means method. First, reliabilities of measures in the literature are quite high (a mean of .88 in Table 1). However, as Grewal et al. (2004, p. 528) conclude: "[e]ven when reliability is fairly high by conventional standards, measurement error can be damaging." One may also overlook that measurement error becomes more severe in latent moderation settings because the reliability of an interaction term is usually lower than the reliability of its components (Busemeyer and Jones 1983; McClelland et al. 2017). Second, researchers might believe that ignoring measurement error leads to underestimated moderation effects and that the means method would therefore be a conservative estimator. However, this is only the case for regressions with a single predictor in which not accounting for measurement error biases parameter estimates to zero (Bollen 1989; Cohen et al. 2003; Grewal et al. 2004; Spearman 1904; Wooldridge 2015). The direction and magnitude of bias in models with multiple predictors, even if some are with and some are without measurement error, are more difficult to predict (Bollen 1989; Cohen et al. 2003; Wooldridge 2015). Third, a comprehensive performance assessment of the six main latent moderation methods is lacking and hinders an informed use of latent moderation methods. This last point motivated this research.

Table 1
Summary of Literature Review

Number of articles	164		
Number of studies	293	Median (SD) sample size across studies	202 (57,493)
Number of moderation effects	656		
		Mean or mode (SD) of data features:	
Number (%) of articles with:		Size of the moderation effect	.17 (.13)
1. Multi-group	4 (2%)	Size of the main effects	.20 (.17)
2. Means	154 (94%)	Correlation X with Z	.17 (.16)
3. Corrected means	1 (1%)	Reliability of Y, X and Z	.88 (.09)
4. Factor scores	7 (4%)	Number of indicators of Y, X and Z	3 (9.33)
5. Product indicators	1 (1%)	Number of scale points of y, x and z	7 (10.83)
6. Latent product	1 (1%)		

Notes: Literature review of moderation analyses in the 2015-2019 volumes of *Journal of Marketing Research* (*JMR*), *Journal of Marketing* (*JM*), *Journal of Consumer Research* (*JCR*) and *Marketing Science* (*Mark. Sci.*). Percentages may not sum to 100% due to rounding and use of multiple methods within an article. Effect sizes are correlations. Effect sizes and correlations report mean and SD (standard deviation) and number of indicators and scale points have modes and SD. Web Appendix A has detailed results.

Our objective is to compare the six methods for latent moderation analysis, both theoretically and empirically, and to provide recommendations for their use. First, we describe the six methods and their differences. Second, we use eight Monte Carlo simulation studies to investigate the statistical properties of the methods under a variety of conditions and in terms of four performance criteria (parameter bias, standard error bias, RMSE and power). The simulations manipulate, respectively, reliability of the measures (Study 1), scale of the indicators (Studies 2a-b), correlation between the (latent) moderation variables (Study 2c), factor loadings (Study 3), and indicator distributions (Study 4a). They show that some methods, specifically the factor scores method and latent product method, outperform the others. In addition, the simulations examine the effects of misspecification, respectively, correlated measurement errors (Study 4b), and ignoring U-shaped (polynomial) effects of the latent variables (Study 4c), and all methods perform worse there. Third, we provide recommendations for future use of the methods and make sample code available to implement the methods.

This paper makes several recommendations for latent moderation analysis. First, when the reliabilities of the moderation variables are close to one, five out of six methods perform well, thus the choice of method is at the researcher's discretion. The corrected means, factor scores, product indicators, and latent product method have parameter bias under 2% and standard error bias under 5% when the reliability of Y, X and Z is a high .95 (Study 1). Under these conditions, the parameter bias of the means method is a slightly higher 8% (and standard error bias is 3%), less than the 10% that is considered acceptable (Feingold 2019; Muthén and Muthén 2002). In contrast, the multi-group method has a bias higher than 20% and should be avoided when moderators have continuous indicators.

Second, our results support the use of the factor scores method and the latent product method in situations where reliabilities of the moderation variables are lower than one. Both

methods perform equally well under most investigated conditions, with bias levels lower than 5%. This is the case when reliabilities are between .75 and .95 (Study 1), for seven-, five- and three-point categorical indicators (Study 2a), correlations between the moderation variables from 0 to .60 (Study 2c) and unequal indicator loadings (Study 3). Researchers might base their choice of either method on the availability in their preferred statistical software.

Third, we identify specific settings in which the multi-group method and product indicators method can be reserved for. The multi-group method can be used for a discrete moderator, although the corrected means, factor scores, product indicators, and latent product method also perform well with biases under 5% (Study 2b). The product indicators method might be chosen over the other methods for non-normally distributed indicators (parameter bias of 5% if skewness of the moderation variables is 3 and excess kurtosis is 10, at a sample size of 200). Yet, its standard error bias can harm statistical conclusion validity (Study 4a).

Web Appendix B overviews sample code to implement all methods in SPSS, Stata, R and Mplus, made available at an OSF repository: https://osf.io/py7jx/?view_only=5d921a6658cf402a80bd1d4996665331.

LATENT MODERATION ANALYSIS

Moderation framework

Assume the following structural latent moderation model:

$$Y = \beta_1 * X + \beta_2 * Z + \beta_3 * XZ + \zeta, \quad (1)$$

where Y is the outcome variable, X is an input variable, Z is a moderator and $\zeta \sim N(0, \sigma_\zeta^2)$ is the structural error term. The parameter β_3 captures the moderation effect and β_1 and β_2 are main effect parameters of respectively X and Z. This paper focuses on latent (unobserved) Y, X and Z but also considers the situation where Z is manifest (observed). We do not consider cases where all Y, X and Z are manifest as standard methods for moderation analysis can be

used in such cases (Cohen et al. 2003; Wooldridge 2015). Without loss of generality, we assume a zero intercept of Y.

The parameters of the latent moderation model cannot be estimated directly because the true scores of Y, X and Z are latent and are reflected in one or more indicator variables that contain measurement error. For exposition, this paper focuses on three indicators per latent variable, the mode in the literature review (Table 1). We consider both continuous and ordered categorical indicators (e.g., Likert scales). The measurement model for X (and analogous for Z and Y) is:

$$x = \Lambda_x * X + \varepsilon_x, \quad (2)$$

where, Λ_x is a vector of loadings or weights and $\varepsilon_x \sim N(0, \theta_x)$ refers to the indicator measurement errors with covariance matrix θ. In terms of notation, we use lowercase (e.g., x) for indicators and uppercase (e.g., X) for latent variables or their approximations with mean or sum scores of indicators (e.g., $\bar{X}$) or factor scores (e.g., $\hat{X}$).

Definitions of key concepts and method performance criteria

This paper is articulated around three key concepts: latent moderation analysis, measurement error and reliability, which Table 2 (Panel A) defines. In addition, Table 2 (Panel B) defines four focal performance criteria to compare the methods for latent moderation analysis: parameter bias, standard error bias, RMSE (Root Mean Squared Error), and power / type I error. Each reflects a statistical property of the estimators that might be affected by measurement error and might vary across methods. This paper mainly focuses on the performance criteria with respect to the moderation effect because it is leading in determining the presence of moderation, but we also consider the main effects as the moderation type (i.e., crossing or not) depends on the sign, size and significance of all three parameters (Cohen et al. 2003).

Parameter bias. Measurement error can bias moderation and main effects. Unbiased estimates are crucial as measures of scientific knowledge and might inform the managerial relevance of effects (Eisend 2015). If Y, X and Z are manifest (and X and Z are normally distributed and uncorrelated), the true moderation effect is (Cohen et al. 2003):

$$\beta_3 = \frac{\mathrm{COV}[Y, XZ]}{\mathrm{VAR}[XZ]}, \tag{3}$$

and analogous for the main effects, where COV refers to a covariance and VAR to a variance. However, suppose that $\bar{X}\bar{Z}$ is a product of scores (e.g., means) of the indicators of X and Z:

$$\bar{X}\bar{Z} = XZ + \varepsilon_{XZ}. \tag{4}$$

where XZ is the true score of the product of X and Z plus normally distributed and random (independent from all true scores and all other εs) measurement error ε_{XZ}. Then $\mathrm{COV}[\bar{Y}, \bar{X}\bar{Z}] = \mathrm{COV}[Y, XZ]$ but $\mathrm{VAR}[\bar{X}\bar{Z}]$is inflated such that the estimated moderation effect $\hat{\beta}_3$ is (Bollen 1989, pp. 154-159):

$$\hat{\beta}_3 = \frac{\mathrm{VAR}[XZ]}{\mathrm{VAR}[XZ] + \mathrm{VAR}[\varepsilon_{XZ}]} * \beta_3 = \rho_{\bar{X}\bar{Z}} * \beta_3, \tag{5}$$

where $\rho_{\bar{X}\bar{Z}}$ is the reliability of $\bar{X}\bar{Z}$, or in other words, the proportion of systematic variance in $\bar{X}\bar{Z}$. Thus, unless $\bar{X}\bar{Z}$ is free of measurement error (i.e., $\rho_{\bar{X}\bar{Z}}$ = 1), the estimated moderation effect is biased towards zero, and the magnitude depends on the reliability of the product. These results are analogous for the main effects if X and Z are uncorrelated, but the direction and the magnitude of bias for all parameters becomes more difficult to determine for correlated predictors. Moreover, bias due to measurement error in variables might carry over to parameter estimates of other variables in the model, even if they do not contain measurement error. Yet, measurement error in Y does not bias moderation effects but might attenuate R^2 (Bollen 1989; Cohen et al. 2003; Wooldridge 2015).

Bias due to measurement error is not specific to latent moderation analysis. Yet it can be more severe in this setting because product terms typically have a lower reliability than

their components.[1] The reliability of a product of $\bar{X}$ and $\bar{Z}$ is (Busemeyer and Jones 1983, Equation 10):

$$\rho_{\bar{X}\bar{Z}} = \frac{\rho_{\bar{X}} * \rho_{\bar{Z}} + r^2_{\bar{X},\bar{Z}}}{1 + r^2_{\bar{X},\bar{Z}}}, \quad (6)$$

where $r^2_{\bar{X},\bar{Z}}$ is the squared correlation between the scores of X and Z. For example, if $\bar{X}$ and $\bar{Z}$ have a reliability of .85, and are correlated .20, the reliability of the product term is a much lower .73. However, a higher correlation between $\bar{X}$ and $\bar{Z}$ increases $\rho_{\bar{X}\bar{Z}}$ and increases the power of the estimated moderation effect (McClelland et al. 2017).

Standard error bias. Measurement error can also bias standard errors (Bollen 1989; Cohen et al. 2003; Van Smeden et al. 2019; Wooldridge 2015). Unbiased standard errors are crucial for valid moderation tests and a valid assessment of the uncertainty of moderation estimates more generally. It is important to note that correcting for measurement error increases standard errors, even if they are unbiased. For correlations, a reasonable approximation for the standard error increase due to the correction is the magnitude that the correlation is biased downward due to measurement error (Hunter and Schmidt 2004, p. 96). However, standard errors are complex functions of the size of the effect, sample size, measure reliabilities, correlations among predictors and the estimated model (Charles 2005; Yuan et al. 2010).

RMSE. The Root Mean Squared Error is based on the sum of the squared bias and the variance of a parameter. It summarizes parameter recovery (lower is better). RMSE can also be used to choose between unbiased estimators. The method with the lowest RMSE (i.e., lowest parameter uncertainty) among unbiased estimators is preferred. Accounting for measurement error decreases parameter bias and thus RMSE. At the same time, the measurement error correction might increase RMSE due to the larger standard error. The net effect on RMSE is difficult to predict.

Table 2
Overview of Key Concepts and Method Performance Criteria

Panel A: Key concepts

Concept	Definition and mathematical illustration
Latent moderation analysis	Definition: moderation analyses in which at least one of the target moderation variables is latent and is measured by one or more reflective indicators that contain measurement error. Mathematical illustration: $$Y = \beta_1 * X + \beta_2 * Z + \beta_3 * XZ + \zeta,$$ where X and/or Z are latent variables that are each reflected in one or more indicators that contain measurement error.
Measurement error	Definition: difference between observed and true values of a score (Wooldridge 2015, p. 288). Mathematical illustration: $$\bar{X}\bar{Z} = XZ + \varepsilon_{XZ},$$ where $\bar{X}\bar{Z}$ is a product of observed (mean) scores, XZ is the product of latent variables X and Z and ε_{XZ} is measurement error.
Reliability	Definition: proportion of systematic variance in a score (Bollen 1989, p. 156). Mathematical illustration: $$\rho_{XZ} = \frac{\mathrm{VAR}[XZ]}{\mathrm{VAR}[XZ] + \mathrm{VAR}[\varepsilon_{XZ}]},$$ where VAR refers to the variance.

Panel B: Method performance criteria

Criterion	Definition and operationalization in Studies 1, 2a-c, 3 and 4a-c	Threshold
Parameter bias	Definition: difference between estimated and true values of $\hat{\beta}$ (Wooldridge 2015, p. 288). Operationalization: $$100 * \mathrm{ABS}\left[\mathrm{MEAN}\left[\sum_{r=1}^{R} \frac{\hat{\beta}_r - \beta}{\beta}\right]\right]$$	≤ 10% (Feingold 2019; Muthén and Muthén 2002)
Standard error bias	Definition: difference between estimated and true values of $\mathrm{SE}[\hat{\beta}]$ (Wooldridge 2015, p. 288). Operationalization: $$100 * \mathrm{ABS}\left[\mathrm{MEAN}\left[\frac{\mathrm{SE}[\hat{\beta}_r] - \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(\hat{\beta}_r - (\frac{1}{R}\sum_{r=1}^{R}\hat{\beta}_r))^2}}{\sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(\hat{\beta}_r - (\frac{1}{R}\sum_{r=1}^{R}\hat{\beta}_r))^2}}\right]\right]$$	≤ 5% (Feingold 2019; Muthén and Muthén 2002)
RMSE (Root Mean Squared Error)	Definition: square root of mean sum of squared bias and variance of $\hat{\beta}$ (Germann et al. 2015). Operationalization: $$\sqrt{\mathrm{MEAN}\left[\left(\hat{\beta}_r - \beta\right)^2 + \mathrm{SE}\left[\hat{\beta}_r\right]^2\right]}$$	Lowest RMSE among unbiased methods (Germann et al. 2015)
Power / Type I error	Definition: probability that $\hat{\beta}$ is found statistically significant at (two-tailed) $p \leq .05$ (Cohen 1988, p. 1). Operationalization: $$100 * \frac{1}{R}\sum_{r=1}^{R} I_r \begin{cases} 1 \text{ if } \mathrm{ABS}\left[\frac{\hat{\beta}_r}{\mathrm{SE}[\hat{\beta}_r]}\right] > 1.96 \\ 0 \text{ otherwise} \end{cases}$$	Power ≥ 80% or type I error ≤ 5% (Cohen 1988; Muthén and Muthén 2002)

Notes: $\hat{\beta}$ refers to an estimated effect for β, the true value of β_1, β_2 or β_3, in Monte Carlo replication r (out of R = 5,000 replications). $ABS[\cdot]$ takes the absolute value, MEAN[·] takes the mean across the R Monte Carlo replications and SE[·] refers to the estimated standard error. Then I is an indicator function and 1.96 is the critical value based on a two-tailed Z-test with 95% confidence.

Power and type I error. Power and type I error are the probability that a parameter of interest is found statistically significant (Cohen 1988, p. 1). High power is crucial to find effects if they truly are non-zero. Measurement error decreases power and thus increases required sample sizes (Grewal et al. 2004). If the true parameter is zero, the analogue to power is type I error. Minimizing it prevents false positive results. RMSE and power complement each other. For instance, a high upward parameter bias can lead to a high power but RMSE would detect that the estimator is problematic. Among unbiased methods, both RMSE and power should provide qualitatively similar results.

SIX METHODS FOR LATENT MODERATION ANALYSIS

Figure 1 visualizes the six methods for latent moderation analysis and provides model equations. Table 3 overviews assumptions of the methods.

Method 1: Multi-group

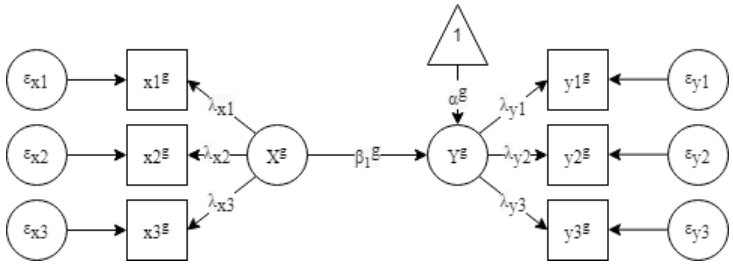
This method estimates separate models for discrete subgroups based on the moderator. We focus on two groups for exposition and as common in moderation analyses (37% of moderation variables in the literature review). The structural model for each group g is:

$$Y^g = \alpha^g + \beta_1^g * X^g + \zeta^g. \quad (7)$$

It does not include an interaction term but estimates a β_1 parameter for each group. The main effect of Z is derived from the intercept α. Constraining β_1 to be equal across groups and testing that model against one with a group-specific β_1 tests moderation. Measurement models as in Equation (2) can be specified for Y and X. Grouping is straightforward for a discrete Z, such as different countries or experimental manipulations and so on. Yet grouping requires discretization based on a median or other split when Z is continuous. Such discretization uses partial information in Z and adds measurement error to the grouping variable (Irwin and McClelland 2001, 2003).

Figure 1
Method Visualizations and Model Equations

Panel A: 1. Multi-group method



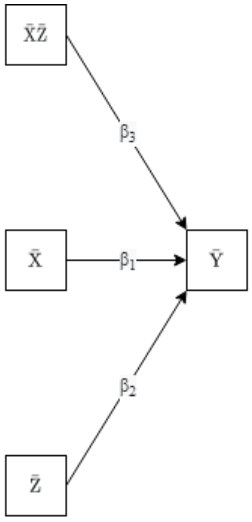
Step 1 – Discretize if Z has continuous indicators:

$$g = \begin{cases} -1 & \text{if } \bar{Z} < \text{MEDIAN}[\bar{Z}], \\ 1 & \text{otherwise.} \end{cases}$$

Step 2 – Specify and estimate the model:

$$y^g = \Lambda_y * Y^g + \varepsilon_y,$$
$$x^g = \Lambda_x * X^g + \varepsilon_x,$$
$$Y^g = \alpha^g + \beta_1^g * X^g + \zeta^g.$$

Panel B: 2. Means method



Step 1 – Take unit weighted means:
Taking means corresponds to the measurement models

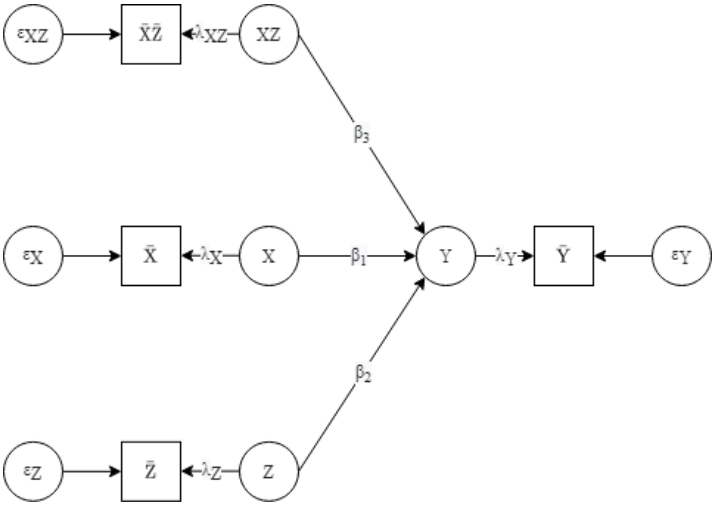
$$y = \Lambda_y * Y + \varepsilon_y,$$
$$x = \Lambda_x * X + \varepsilon_x,$$
$$z = \Lambda_z * Z + \varepsilon_z,$$

where for each measurement model, the elements in Λ and the elements in ε are constrained to be equal (McNeish and Wolf 2020).

Step 2 – Specify and estimate the structural model:

$$\bar{Y} = \beta_1 * \bar{X} + \beta_2 * \bar{Z} + \beta_3 * \bar{X}\bar{Z} + \zeta.$$

Panel C: 3. Corrected means method



Step 1 – Take unit weighted means (as in "2. Means") and estimate reliability ρ (e.g., with Cronbach's alpha).

Step 2 – Specify and estimate the model:

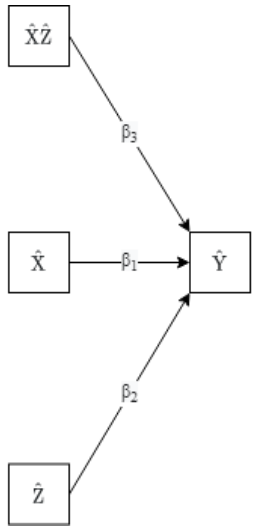
$$\bar{Y} = \lambda_Y * Y + \varepsilon_Y,$$
$$\bar{X} = \lambda_X * X + \varepsilon_X,$$
$$\bar{Z} = \lambda_Z * Z + \varepsilon_Z,$$
$$\bar{X}\bar{Z} = \lambda_{XZ} * XZ + \varepsilon_{XZ},$$
$$Y = \beta_1 * X + \beta_2 * Z + \beta_3 * XZ + \zeta.$$

For identification, fix the λs to 1 and the εs fix the amount of measurement error:

$$VAR[\varepsilon_Y] = (1 - \rho_{\bar{Y}}) * \text{VAR}[\bar{Y}],$$
$$VAR[\varepsilon_X] = (1 - \rho_{\bar{X}}) * \text{VAR}[\bar{X}],$$
$$VAR[\varepsilon_Z] = (1 - \rho_{\bar{Z}}) * \text{VAR}[\bar{Z}],$$
$$VAR[\varepsilon_{XZ}] = (1 - \rho_{\bar{X}\bar{Z}}) * \text{VAR}[\bar{X}\bar{Z}].$$

Figure 1 (CONTINUED)

Panel D: 4. Factor scores method



Step 1 – Specify and estimate measurement models:

First, $y = \Lambda_y * Y + \varepsilon_y$ (1-CFA). And then

$x = \Lambda_x * X + \varepsilon_x$ simultaneously with

$z = \Lambda_z * Z + \varepsilon_z$, correlating X and Z (2-CFA)

Extract Bartlett scores for Y and regression scores for X and Z:

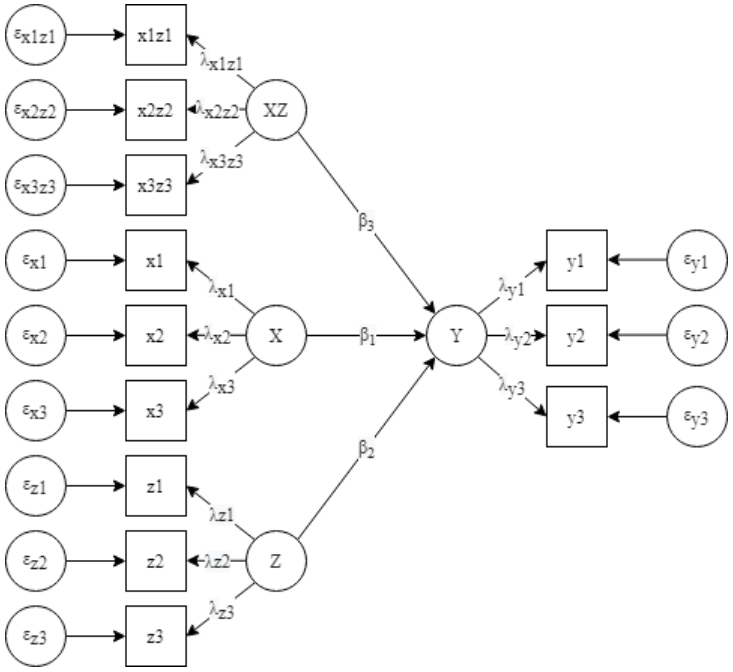
$$\hat{F}_{Bartlett} = D\Theta^{-2}\Lambda(\Lambda^T\Theta^{-2}\Lambda)^{-1},$$

$$\hat{F}_{Regression} = D\Sigma_{(o)}^{-1}\Lambda\Phi.$$

Step 2: Specify and estimate the structural model:

$$\hat{Y} = \beta_1 * \hat{X} + \beta_2 * \hat{Z} + \beta_3 * \hat{X}\hat{Z} + \zeta.$$

Panel E: 5. Product indicators method

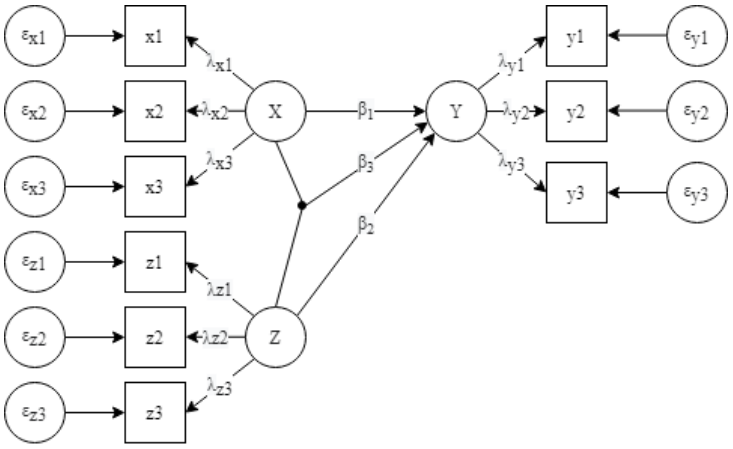


Specify and estimate the model:

$$y = \Lambda_y * Y + \varepsilon_y,$$
$$x = \Lambda_x * X + \varepsilon_x,$$
$$z = \Lambda_z * Z + \varepsilon_z,$$
$$xz = \Lambda_{xz} * XZ + \varepsilon_{xz},$$
$$Y = \beta_1 * X + \beta_2 * Z + \beta_3 * XZ + \zeta,$$

where xz are products of x and z.

Panel F: 6. Latent product method



Specify and estimate the model:

$$y = \Lambda_y * Y + \varepsilon_y,$$
$$x = \Lambda_x * X + \varepsilon_x,$$
$$z = \Lambda_z * Z + \varepsilon_z,$$
$$Y = \beta_1 * X + \beta_2 * Z + \beta_3 * XZ + \zeta.$$

Notes: The "steps" denote whether the measurement and structural models are estimated separately (in two steps) or not. All visualizations have three indicators for Y, X and Z for exposition. Circles are latent variables, and boxes are manifest indicators. Unidirectional arrows refer to loadings λ and regression paths β. Then ζs are structural error terms, ommitted from visualizations for exposition, and εs are measurement errors. Error variances, latent variances and covariances between explanatory variables X, Z, and XZ are omitted for brevity. Superscript g refers to a discrete grouping variable and the triangle "1" is an intercept α (Panel A), bars (e.g., $\bar{X}$) denote means (Panels B and C), and hats (e.g., $\hat{X}$) denote estimated factor scores (Panel D). Panel E uses the "matched pairs" strategy to form three product indicators, but readily extends to other indicator pairings. In Panel F, the dot connecting X and Z refers to the moderation effect being inferred from the joint distribution of the indicators of X and Z and not based on observed product terms of X and Z and/or their indicators (Muthén and Muthén 2019).

Table 3
Overview of Method Assumptions

Assumption	1. Multi-group	2. Means	3. Corrected means	4. Factor scores	5. Product indicators	6. Latent product
Measurement model						
Indicator distribution						
$x \sim MVN(\mu_x, \Sigma_x)$	Yes	-	Yes	Yes	Yes	Yes
$z \sim MVN(\mu_z, \Sigma_z)$	No, discrete	-	Yes	Yes	Yes	Yes
$xz \sim MVN(\mu_{xz}, \Sigma_{xz})$	-	-	-	-	Yes	-
Account for implied non-normality in y	No	No	No	No	No	Yes
Indicator measurement errors						
All $\varepsilon_x \sim MVN(0, \Theta_x)$ freely estimated	Yes	No, fixed and equal	No, fixed and equal but accounted for	Yes	Yes	Yes
All $\varepsilon_z \sim MVN(0, \Theta_z)$ freely estimated	No, fixed and equal	No, fixed and equal	No, fixed and equal but accounted for	Yes	Yes	Yes
All $\varepsilon_{xz} \sim MVN(0, \Theta_{xz})$ freely estimated	-	-	-	-	Yes[a]	-
All $\varepsilon_y \sim MVN(0, \Theta_y)$ freely estimated	Yes	No, fixed and equal	No, fixed and equal but accounted for	Yes	Yes	Yes
Indicator loadings						
All Λ_x freely estimated	Yes	No, fixed and equal	No, fixed and equal	Yes	Yes	Yes
All Λ_z freely estimated	No, fixed and equal	No, fixed and equal	No, fixed and equal	Yes	Yes	Yes
All Λ_{xz} freely estimated	-	-	-	-	Yes[a]	-
All Λ_y freely estimated	Yes	No, fixed and equal	No, fixed and equal	Yes	Yes	Yes
Structural model						
$\zeta \sim N(0, \sigma_\zeta^2)$, uncorrelated with y, x, z, X, Z and all Θ	Yes	Yes	Yes	Yes	Yes	Yes

[a]: The product indicators method freely estimates the loadings and measurement errors of the product indicators but using "matched pairs" assumes that all product indicators are equally good representatives of the latent interaction factor XZ because the moderation result might depend on the choice of indicator pairs if indicators are not equally good, which is undesirable (Foldnes and Hagtvet 2014; Marsh et al. 2004).

Notes: All methods except the latent product method use standard maximum likelihood estimation, which uses the expectation maximization (EM) algorithm that converges to maximum likelihood estimates (Dempster et al. 1977; Klein and Moosbrugger 2000). MVN(·) is the multivariate normal distribution and N(·) is the normal distribution. The "-" denotes that the assumption is not applicable, that is, the means method does not directly use a measurement model so it does not assume a distribution of the indicators. Similarly, the multi-group and latent product methods do not use manifest interactions or product terms to estimate the moderation effect; only the product indicators method uses products of indicators in the measurement model.

Method 2: Means

This method uses unit weighted mean (or sum) scores of the indicators. Although mean scores can be used without estimating a measurement model, McNeish and Wolf (2020) show that unit weighted means are analogous to assuming a parallel measurement model that constrains indicators to be equally weighted with equal measurement error variances. The structural model then uses the mean scores to estimate the moderation effect without accounting for measurement error in the mean scores:

$$\bar{Y} = \beta_1 * \bar{X} + \beta_2 * \bar{Z} + \beta_3 * \bar{X}\bar{Z} + \zeta. \quad (8)$$

The means $\bar{X}$ and $\bar{Z}$ can be mean-centered prior to computing the interaction term $\bar{X}\bar{Z}$ to facilitate interpretation and reduce unessential multicollinearity (Cohen et al. 2003; Irwin and McClelland 2001).

Method 3: Corrected means

This method uses a product of mean scores, as the means method does, but accounts for measurement error in the scores by using reliability estimates. A measurement model as in Equation (2) can be used but with loadings and measurement errors fixed for identification (Bollen 1989). For example, for XZ, the loading is $\lambda_{XZ} = 1$ and the error variance is $\sigma^2_{\varepsilon_{XZ}} = (1 - \rho_{\bar{X}\bar{Z}}) * \sigma^2_{\bar{X}\bar{Z}}$, where $\sigma^2_{\bar{X}\bar{Z}}$ is the variance of $\bar{X}\bar{Z}$, and $\rho_{\bar{X}\bar{Z}}$ is its reliability. Reliabilities of Y, X and Z can be estimated with estimators such as Cronbach's alpha, assuming unit weighted indicators. Then $\rho_{\bar{X}\bar{Z}}$ can be estimated with Equation (6). The structural model relates the latent variables as in Equation (1). Statistically, the mean scores across multiple indicators are single-indicators of the latent variables. Thus, the corrected means method can also be used for single-indicator measures if their reliability can be estimated (e.g., Pieters 2017, pp. 699-700).

Method 4: Factor scores

This method uses factor scores that estimate the latent variable scores with linear combinations of the indicators. A first step extracts factor scores from measurement models as in Equation (2) that freely estimate measurement errors and loadings. A second step regresses factor scores of Y on those of X, Z and the product:

$$\hat{Y} = \beta_1 * \hat{X} + \beta_2 * \hat{Z} + \beta_3 * \hat{X}\hat{Z} + \zeta. \quad (9)$$

There are multiple ways to estimate factor scores. In the context of non-moderation models, Skrondal and Laake (2001) and Devlieger et al. (2016) have shown that using Bartlett factor scores for Y and regression factor scores for predictors produces estimates without bias:

$$\hat{F}_{Bartlett} = D\Theta^{-2}\Lambda(\Lambda^T\Theta^{-2}\Lambda)^{-1}, \quad (10)$$

$$\hat{F}_{Regression} = D\Sigma_{(o)}^{-1}\Lambda\Phi, \quad (11)$$

where D is a matrix of indicator-level data, Θ is the variance covariance matrix of the indicator measurement errors, Λ is the matrix of estimated loadings, $\Sigma_{(o)}^{-1}$ is the observed covariance matrix of the indicators, and Φ is the variance covariance matrix of the latent variables (Lastovicka and Thamodaran 1991). Bartlett factor scores account for measurement error in Y and regression factor scores account for measurement error in the predictors; combining these factor scores recovers the parameters in non-moderation models without parameter bias (Devlieger et al. 2016; Skrondal and Laake 2001). We apply this to the context of latent moderation.

There are several ways to specify the measurement models. Measurement models for Y, X and Z can be estimated jointly or separately with confirmatory (CFA) or exploratory (EFA) factor analyses estimated with maximum likelihood. Skrondal and Laake (2001) have shown that separate factor analyses for Y (1-CFA or unrotated 1-EFA) and the predictors are necessary to avoid parameter bias. The predictors need to be combined in a joint

confirmatory factor analysis (2-CFA of X and Z) because the 2-CFA accounts for the factor correlation of X with Z. Skrondal and Laake (2001, pp. 572-573) then show with analytical proofs that estimates are unbiased if the correlation is accounted for. Web Appendix C has additional details, including the extension to (moderated) mediation models.

Method 5: Product indicators

This method specifies a measurement model analogous to Equation (2) for products of indicators, while simultaneously estimating the structural model of Equation (1). There are several ways to specify this model. They differ in the product indicators to pair for moderation analysis and the used constraints to estimate the model. Early on, Kenny and Judd (1984) proposed using a measurement model of product indicators that required multiple constraints on the indicator loadings and measurement error variances. Foldnes and Hagtvet (2014) showed based on simulation studies and real-world data that there might be considerable variation in moderation estimates depending on the method to pair indicators. Using a single pair of indicators uses limited information (Jöreskog and Yang 1996), whereas using all pairs of indicators uses all information but might lead to overly complex models (Marsh et al. 2004). Marsh et al. (2004) proposed a compromise "matched pairs" approach, using all indicators of X and Z but each indicator only once. This approach trades off the use of all indicators while limiting model complexity—avoiding correlated measurement errors for pairs that have common components—with acceptable bias and variance implications (Marsh et al. 2004). Lin et al. (2010) have shown that using matched pairs and "double mean-centering" the indicator pairs works well. It avoids the need for constraints, other than those for identification, on the indicator loadings and measurement error variances.

A Web of Science citation analysis signals that the product indicators method is uncommonly used in the focal journals of the literature review, even beyond the included volumes. The three citations of Marsh et al. (2004) apply the method whereas three out of

four citations of Kenny and Judd (1984) refer to its methodological contribution without application. Moreover, the matched pairs approach in Marsh et al. (2004) has been accumulating more total citations (596) within and outside the marketing domain than other approaches have (588 citations of Kenny and Judd (1984), 272 of Jöreskog and Yang (1996), 70 of Lin et al. (2010) and 13 of Foldnes and Hagtvet (2014)).

Based on this, we use the matched pairs approach with double mean-centering to represent the product indicators method here. In our running example where both X and Z have three indicators taking matched pairs results in three product indicators of mean-centered variables, for example x1z1, x2z2 and x3z3 (Marsh et al. 2004), that are subsequently mean-centered once more (Lin et al. 2010).

Method 6: Latent product

This method estimates the moderation effect from the latent product of X and Z (Klein and Moosbrugger 2000). The latent product method is motivated by the non-normality in Y that is due to the moderation specification (Klein and Moosbrugger 2000). Products of variables (e.g., XZ) are usually non-normally distributed, even if their components (here: X and Z) are normally distributed. Because Y is a function of the non-normally distributed XZ if there is a non-zero moderation effect, it is also non-normally distributed (Moosbrugger et al. 1997). Web Appendix D further details this and provides an illustrative example. The latent product method takes the non-normality in Y directly into account. It is therefore based on an analysis of the indicator distribution and uses the raw data for estimation, unlike the other methods for which the observed covariance matrix is sufficient. The non-normal indicator distribution can be approximated by a weighted sum or finite mixture of normal distributions (Klein and Moosbrugger 2000). The mixture distribution then becomes a tool to estimate the moderation effect from the latent product of the latent X and Z. Web Appendix E details this.

Commonalities and differences between the methods

In terms of commonalities between the six methods, they all rely on the same estimation approach. All methods except for the latent product method use standard maximum likelihood estimation (Bollen 1989). The latent product method uses an expectation maximization (EM) algorithm that converges to maximum likelihood estimates too (Klein and Moosbrugger 2000), even though EM can be computationally intensive and more sensitive to local maxima of the likelihood (Dempster et al. 1977).

The structural moderation models of five out of six methods (all except the multi-group method) are virtually identical. The crucial difference is in the specification and assumptions of the measurement model (Table 3). The means method takes unit weighted mean scores of the indicators that assume a parallel measurement model (McNeish and Wolf 2020). The means method does not account for the remaining measurement error in the scores. The corrected means method accounts for this shortcoming of the means method by fixing the amount of measurement error in the variables based on reliability estimates. Yet, it maintains the assumptions of a parallel measurement model. The equal indicator weighting biases reliability estimates downward and therefore might lead to upward parameter bias in the moderation effect even if measurement error is accounted for (McNeish and Wolf 2020).

Whereas the means method and corrected means method assume equally weighted indicators, the measurement models of the factor scores method, product indicators method and latent product method freely estimate the loadings and measurement error variances. There are three differences between these methods. First, the factor scores method is a two-step approach that separately estimates measurement and structural models, whereas the product indicators method and latent product method estimate the measurement model and moderation effect simultaneously. Second, although the factor scores method and latent product method use a product of latent variables or their scores in the case of the factor scores

method, the product indicators method uses products of matched pairs of indicators. This assumes that the product indicators are representatives of all possible pairs, essentially assuming equally weighted indicators. There is considerable variation in moderation estimates as a result of different indicator pairings if indicators are not equally good, which is undesirable (Foldnes and Hagtvet 2014; Marsh et al. 2004). Third, the latent product method is the only approach that accounts for the non-normally distributed indicators of Y due to the interaction (Klein and Moosbrugger 2000). However, it maintains the assumption of normally distributed indicators of X and Z, as well do the factor scores method and product indicators method. Yet interestingly, the product indicators method uses products of indicators that rarely meet the assumption of them being normally distributed because products are usually non-normally distributed even if their components are normally distributed (Moosbrugger et al. 1997; Oliveira et al. 2016). Web Appendix D details this.

The multi-group method can include measurement models for the indicators of Y and X to account for indicator measurement error but does not rely on a product of variables and estimates models for discrete subgroups based on the moderators. Although naturally discrete moderators—such as different countries, owners of different brands, genders, experimental manipulations and so on—can readily be used as grouping variables, grouping by discretizing continuous moderators adds measurement error to the grouping variable and can lead to parameter bias and a decrease of power (Irwin and McClelland 2001, 2003).

In sum, the six methods for latent moderation analysis are all based on maximum likelihood estimation but the main differences are in their approach and assumptions of the measurement model.

Table 4
Summary of Study Designs

Study	Methods compared	Reliability of Y, X and Z	Indicator scale of y, x and z	Correlation of X with Z	Indicator loadings	Distribution of x and z	Indicator measurement errors	Structural model specification
Study 1: Reliability of measures	1-6	.95, .85, .75	Continuous	.20	Equal	Normal	Uncorrelated	Correctly specified
Study 2a: Ordered categorical indicators	1-6	.85	Ordered categorical (7-, 5-, 3-point scales)	.20	Equal	Ordered categorical	Uncorrelated	Correctly specified
Study 2b: Discrete moderator	1-6	.95, .85, .75	y and x: continuous z: discrete (binary)	0	Equal	x: normal z: discrete (binary)	Uncorrelated	Correctly specified
Study 2c: Correlation of X with Z	1-6	.85	Continuous	0, .20, .40, .60	Equal	Normal	Uncorrelated	Correctly specified
Study 3: Unequal indicator loadings	4-6	.85	Continuous	.20	Unequal (1, 1.5, .50)	Normal	Uncorrelated	Correctly specified
Study 4a: Non-normally distributed indicators	4-6	.85	Continuous	.20	Equal	Non-normal[a]	Uncorrelated	Correctly specified
Study 4b: Correlated measurement errors	4-6	.85	Continuous	.20	Equal	Normal	Correlated .50 (x with y, x with z, x with x)[b]	Correctly specified
Study 4c: Structural model is misspecified	4-6	.85	Continuous	0, .20, .40, .60	Equal	Normal	Uncorrelated	Misspecified[c]

[a]: Study 4a has non-normality in x and z due to non-normality in X and Z (skewness/excess kurtosis of X and Z is 1/2 or 3/10).
[b]: Study 4b condition 'x with x' means that indicators of X are .50 intercorrelated (with other x-indicators).
[c]: Study 4c generates a polynomial of X ($Y = \beta_1 X + \beta_2 Z + \beta_4 X^2$) and estimates it with Equation (1).

Notes: Shading denotes the focus of each study. Studies 1 and 2a-c investigate all six methods. Studies 3 and 4a-c drop the multi-group and means methods due to their performance in Studies 1 and 2a-c. All studies vary the sample size: 100, 150, 200, 300, 500, 750 and 1,500.

OVERVIEW OF MONTE CARLO SIMULATION STUDIES

We conduct Monte Carlo simulations to compare the statistical properties of the latent moderation methods across conditions. We use simulations because method performance and impact of design factor on method performance are difficult to derive analytically (Muthén and Muthén 2002; Skrondal 2000).

Summary of studies

Table 4 summarizes the designs of eight Monte Carlo simulation studies (1, 2a-c, 3, and 4a-c) that focus on a variety of conditions. All studies, unless indicated otherwise, are under the following conditions. They generate standard normally distributed Y, X and Z. Data generation is based on values from the literature review as much as possible, thus mimicking real-world situations (Table 1). The latent Y, X and Z variables have three indicators, which is most common in the literature review, that are equally good. Reliabilities of Y, X and Z are .85, which is about the mean in our literature review and the mean in a recent review of mediation analyses (Pieters 2017). The moderation and main effect sizes are .20, which are about the mean values in the literature review and small-to-medium effects (Cohen 1988). The correlation between X and Z is .20, about the mean in the literature review. Sample sizes are 100, 150, 200 (median in the literature review), 300, 500, 750 or 1,500. About 80% of the studies in the literature review have sample sizes between 100 and 1,500.

For each study, we generate 5,000 replications (datasets) per cell in R (R Core Team 2020) using common random number seeds to increase precision and for reproducibility (Skrondal 2000). R package lavaan (Rosseel 2012) implements all methods except for the latent product method. For that we call Mplus 8.3 (Muthén and Muthén 2019) from R via MplusAutomation (Hallquist and Wiley 2018). The OSF repository at https://osf.io/py7jx/?view_only=5d921a6658cf402a80bd1d4996665331 has simulation code for all studies.

Method performance criteria

Table 2 (Panel B) has the operationalizations of the performance criteria to compare the methods. We calculate parameter bias by taking the deviations of the estimated main or moderation effect parameter $\hat{\beta}$ from its true value β and dividing by the true value such that the bias is on a percentage scale. We then take the mean across Monte Carlo replications. Similarly, standard error bias is the mean deviation of the estimated standard error from the true standard error, of which the standard deviation of the estimated parameter across replications is an estimate (Muthén and Muthén 2002). RMSE takes the square root of the sum of the parameter bias and estimated variance (squared standard error) and an estimate of power (or type I error if the true parameter is zero) is the percentage of Monte Carlo replications for which the parameter of interest is statistically significant at two-tailed $p \leq .05$.

We evaluate the methods as follows. We first calculate biases in parameters and standard errors and retain the unbiased methods. Common acceptable levels of absolute parameter bias are ≤ 10% and ≤ 5% for standard error bias (Feingold 2019; Muthén and Muthén 2002). For the methods that meet these criteria, we consider RMSE and power. However, these criteria are not interpretable for biased methods because downward standard error bias can lead to low RMSE and upward parameter bias can lead to high power. Common thresholds are ≥ 80% for power and ≤ 5% type I error (Cohen 1988; Muthén and Muthén 2002). Panel B in Table 2 summarizes these thresholds.

Table 5 summarizes the performance criteria for all methods across the conditions for each study at about the median sample size of 200 in the literature review. Web Appendices F-M and the material on OSF plot detailed results for the moderation and main effects.

Table 5
Multi-Group (M1), Means (M2), Corrected Means (M3), Factor Scores (M4), Product Indicators (M5) and Latent Product (M6) Method Performance

	Performance criterion of latent moderation method M1-6 at a sample size of 200																							
	Parameter bias (in %)						Standard error bias (in %)						RMSE						Power / Type I error (in %)					
Study and condition	M1	M2	M3	M4	M5	M6	M1	M2	M3	M4	M5	M6	M1	M2	M3	M4	M5	M6	M1	M2	M3	M4	M5	M6
Study 1: Reliability of measures																								
Reliability of Y, X and Z is .95	37	8	2	1	2	1	4	3	4	1	4	3	.34	.29	.31	.30	.31	.31	60	80	80	80	80	80
Reliability of Y, X and Z is .85	40	26	3	2	4	2	4	3	4	2	10	3	.39	.31	.37	.35	.39	.37	46	66	65	65	61	64
Reliability of Y, X and Z is .75	43	41	6	3	13	4	5	3	4	2	22	4	.45	.34	.44	.41	1.1	.45	34	52	50	51	37	48
Study 2a: Ordered categorical indicators																								
7-point ordered categorical scales	34	27	1	3	1	2	4	3	3	1	5	2	.40	.32	.38	.36	.39	.38	42	60	59	60	58	58
5-point ordered categorical scales	54	29	1	4	1	2	6	2	3	1	5	1	.41	.32	.39	.37	.40	.39	40	58	57	58	54	56
3-point ordered categorical scales	73	38	36	5	8	4	13	8	43	2	21	3	.46	.34	.51	.40	.47	.43	33	49	46	48	43	45
Study 2b: Discrete moderator																								
Reliability of Y, X and Z is .95	1	5	<1	1	1	<1	2	3	4	1	4	1	.30	.29	.30	.29	.30	.30	81	81	81	81	81	81
Reliability of Y, X and Z is .85	1	15	1	2	1	1	3	3	3	<1	4	3	.35	.31	.33	.33	.34	.34	70	71	70	71	70	70
Reliability of Y, X and Z is .75	2	25	2	3	2	2	5	3	3	1	5	4	.40	.33	.38	.37	.39	.39	58	60	59	59	58	58
Study 2c: Correlation of X with Z																								
Correlation X with Z is 0	27	27	3	2	4	1	4	3	3	1	9	2	.37	.31	.37	.35	.39	.37	44	64	63	63	59	62
Correlation X with Z is .20	40	26	3	2	4	2	4	3	4	2	10	3	.39	.31	.37	.35	.39	.38	46	66	65	65	61	64
Correlation X with Z is .40	58	24	2	2	4	2	5	3	3	2	10	3	.43	.31	.38	.36	.40	.38	50	72	71	71	68	71
Correlation X with Z is .60	86	21	2	2	3	1	5	4	4	2	10	3	.50	.31	.41	.40	.43	.42	57	79	79	80	76	79
Study 3: Unequal indicator loadings																								
Indicator loadings are unequal	-	-	25	3	4	4	-	-	4	3	12	5	-	-	.43	.34	.38	.37	-	-	67	69	67	68

Table 5 (CONTINUED)

	Performance criterion of latent moderation method M1-6 at a sample size of 200																							
	Parameter bias (in %)						Standard error bias (in %)						RMSE						Power / Type I error (in %)					
Study and condition	M1	M2	M3	M4	M5	M6	M1	M2	M3	M4	M5	M6	M1	M2	M3	M4	M5	M6	M1	M2	M3	M4	M5	M6
Study 4a: Non-normally distributed indicators																								
x and z are moderately non-normally distributed (X and Z skewness is 1, excess kurtosis is 2)	-	-	5	3	5	5	-	-	6	5	16	4	-	-	.38	.35	.41	.38	-	-	66	69	62	68
x and z are severely non-normally distributed (X and Z skewness is 3, excess kurtosis is 10)	-	-	19	14	5	14	-	-	12	13	32	7	-	-	.40	.34	.43	.37	-	-	79	83	76	82
Study 4b: Correlated measurement errors																								
Measurement errors of x are correlated .50 with those of y	-	-	48	45	48	47	-	-	3	2	10	2	-	-	.39	.38	.42	.40	-	-	66	66	63	65
Measurement errors of x are correlated .50 with those of z	-	-	6	8	7	6	-	-	4	2	9	2	-	-	.37	.36	.40	.38	-	-	56	64	60	63
Measurement errors of x are intercorrelated .50	-	-	21	12	21	21	-	-	3	2	6	2	-	-	.34	.35	.35	.35	-	-	60	60	57	59
Study 4c: Structural model is misspecified [a]																								
Correlation X with Z is 0	-	-	1	2	1	1	-	-	10	11	13	14	-	-	.38	.37	.39	.39	-	-	8	8	8	9
Correlation X with Z is .20	-	-	8	9	8	9	-	-	10	10	14	13	-	-	.40	.39	.42	.41	-	-	18	21	17	22
Correlation X with Z is .40	-	-	14	15	14	15	-	-	9	7	12	9	-	-	.45	.44	.46	.47	-	-	43	48	40	48
Correlation X with Z is .60	-	-	18	18	18	19	-	-	7	5	11	7	-	-	.52	.51	.53	.54	-	-	68	73	65	72

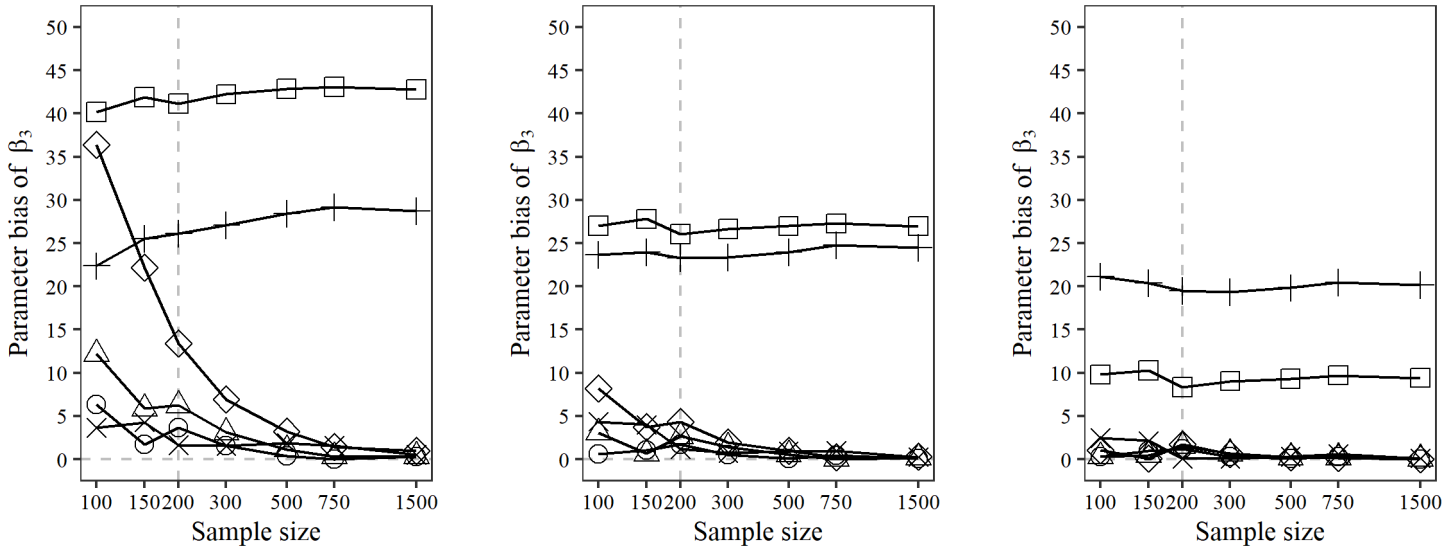
[a]: Study 4c generates a polynomial of X ($Y = \beta_1 X + \beta_2 Z + \beta_4 X^2$) and estimates it with Equation (1).

Notes: Shaded cells indicate acceptable levels of parameter bias (maximum of the moderation and main effects) ≤ 10% and standard error bias ≤ 5% (Feingold 2019; Muthén and Muthén 2002) at a sample size of 200, which is about the median in the literature review (Table 1). Reported RMSE sums RMSE of the moderation and main effects and reported power is the estimated power of β_3 as target moderation test. Then "-" indicates that the multi-group and means methods were excluded based on Studies 1 and 2a-c. Method labels are multi-group (M1), means (M2), corrected means (M3), factor scores (M4), product indicators (M5), latent product (M6).

Figure 2
Study 1: Performance Criteria for the Moderation Effect (β_3)

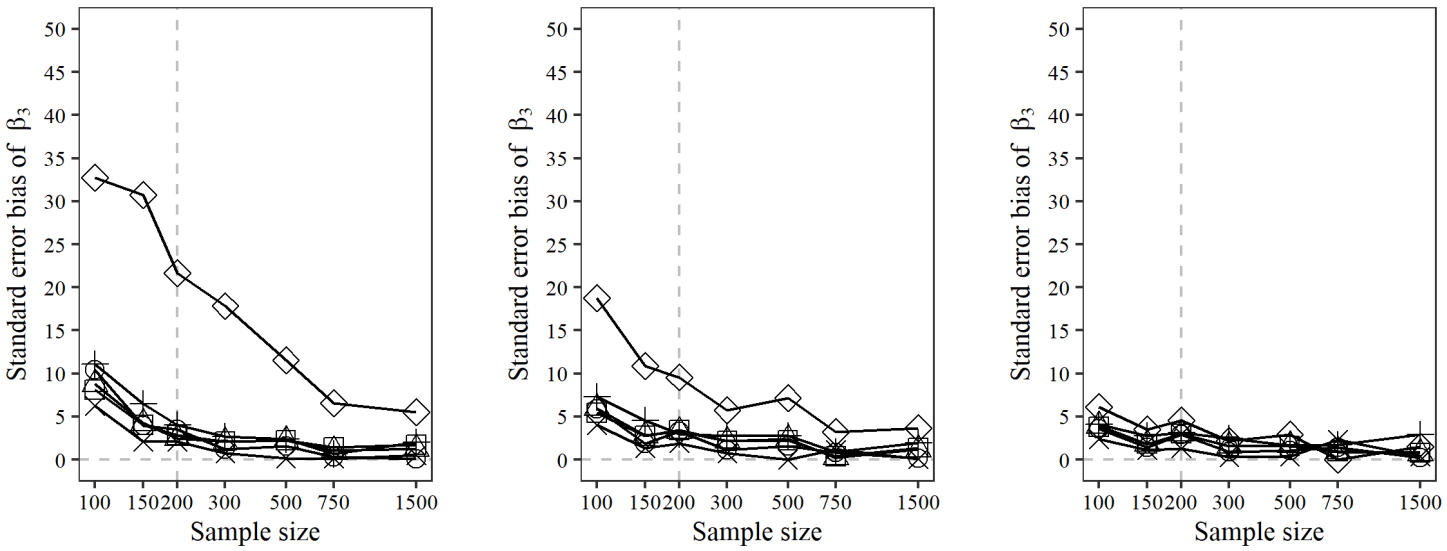
Panel A: Parameter bias of β_3

Reliability of Y, X and Z is .75 | Reliability of Y, X and Z is .85 | Reliability of Y, X and Z is .95



Panel B: Standard error bias of β_3

Reliability of Y, X and Z is .75 | Reliability of Y, X and Z is .85 | Reliability of Y, X and Z is .95



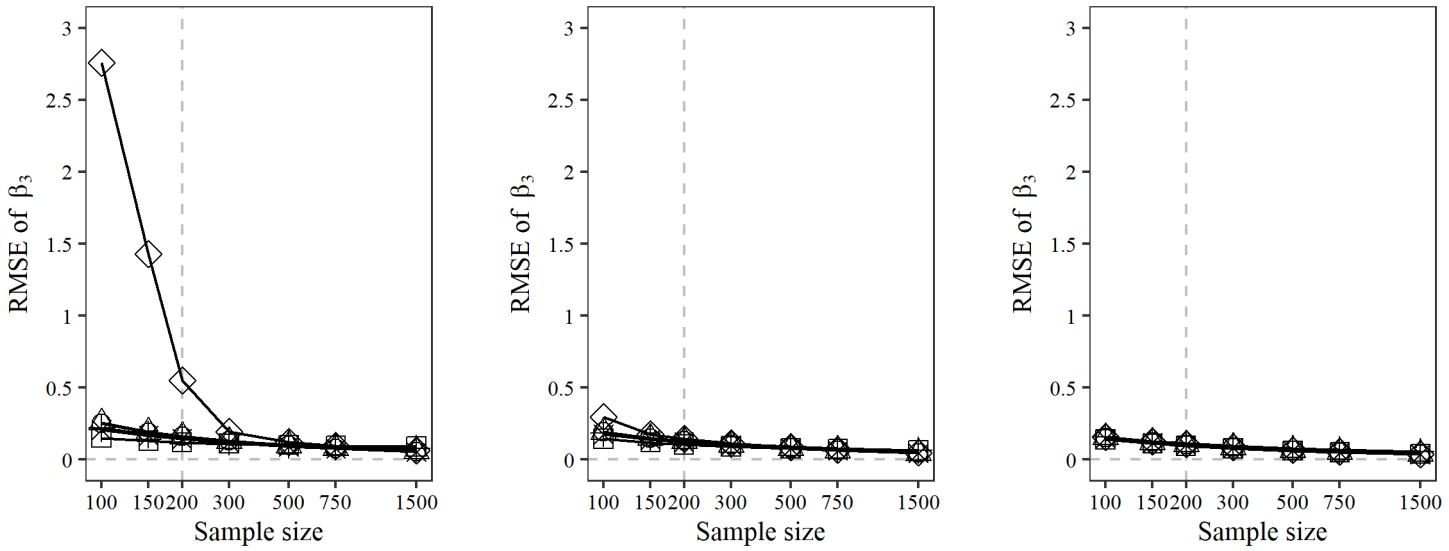
Legend:

1. Multi-group
2. Means
3. Corrected means
4. Factor scores
5. Product indicators
6. Latent product

Figure 2 (CONTINUED)

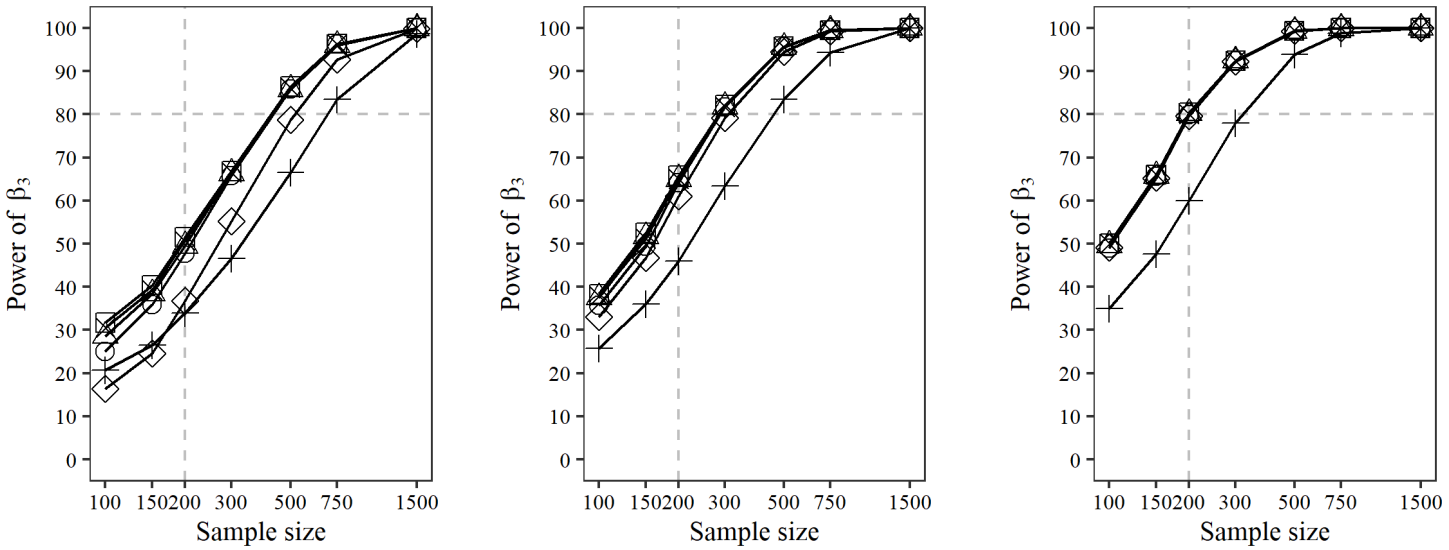
Panel C: RMSE of β_3

Reliability of Y, X and Z is .75 | Reliability of Y, X and Z is .85 | Reliability of Y, X and Z is .95

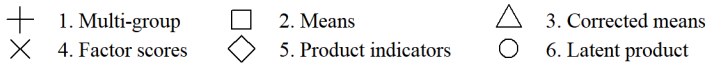


Panel D: Power of β_3

Reliability of Y, X and Z is .75 | Reliability of Y, X and Z is .85 | Reliability of Y, X and Z is .95



Legend:



Notes: Plots visualize method parameter bias, standard error bias, root mean squared error (RMSE) and power (as defined in Table 2) of the moderation effect (β_3) across sample sizes (log scale) and reliabilities of Y, X and Z. Horizontal dashed lines indicate parameter bias, standard error bias and RMSE of zero and power of 80%. Vertical dashed lines indicate a sample size of 200, which is about the median in the literature review (Table 1).

STUDY 1: RELIABILITY OF MEASURES

Design

Study 1 focuses on measure reliability as a determinant of method performance. The design is: 6 (Method) × 7 (Sample size) × 3 (Reliability of Y, X and Z: .95, .85 or .75). The reliability levels .95, .85 and .75 are approximately the mean in the literature review, plus and minus one standard deviation (Table 1). These levels are respectively excellent, good, and acceptable reliability (Peterson 1994). We expect that the multi-group method is biased, has a high RMSE and low statistical power because discretizing the continuous indicators of the moderator adds measurement error. We expect the means method to be biased, but the bias to decrease when the reliability increases. In contrast, the latent product method should recover parameters well. An open question is whether the corrected means method, the factor scores method and the product indicators method perform similar to the latent product method. Moreover, it is unclear how these methods perform at lower measure reliabilities (i.e., .75) and/or in smaller samples (e.g., 100 observations).

Results

Panels A-D in Figure 2 plot performance of the moderation effect estimates (y-axis) across sample sizes (x-axis) for each method (symbols) and across measure reliability levels (.75 in left plot, .85 in center plot and .95 in right plot of each panel). Overall, methods perform better and more similar to each other when measure reliability and sample size increase. However, there are several key performance differences between methods.

Parameter bias (Panel A). The multi-group method is biased, even at high reliability levels of .95 and large sample sizes (e.g., 1,500) with a bias of about 20%. Similarly, the means method is biased for respectively 41% and 26% at reliabilities of .75 and .85. Increasing sample size does not reduce bias, making the multi-group method and means method inconsistent estimators (Wooldridge 2015, p. 287). Yet, the bias of the means method

at a reliability of .95 is 8%, which can be acceptable (Table 2). At that reliability, differences between methods become smaller. The corrected means, factor scores, product indicators, and the latent product method have biases about 1-2%. Differences between methods become larger at lower reliabilities. The product indicators method is unbiased only at larger sample sizes (e.g., ≥ 300) at a reliability of .75. Overall, the corrected means method, factor scores method and latent product method are unbiased (parameter bias below 6% across reliabilities and at a sample size of 200).

Standard error bias (Panel B). All methods except the product indicators method have standard error biases under 5% for sample sizes of at least 200 observations. The product indicators method has biased standard errors (up to about 33% at a reliability of .75) when measure reliability is smaller than .95. Its standard error bias reduces when sample size increases (e.g., bias about 5% at a reliability of .75 and sample size of 1,500).

RMSE (Panel C). RMSE differences are small among the unbiased methods (e.g., between .12 (factor scores method) and .14 (product indicators method) at a reliability of .85 and sample size of 200). The means method offers the best RMSE in smaller samples (≤ 500 observations). However, it is biased and should therefore not be used. The product indicators method has a high RMSE, .55 at a reliability of .75 and a sample size of 200, due to its upward standard error bias.

Power (Panel D). Among unbiased methods, the factor scores method has the highest power: an estimated 65% at measure reliabilities of .85 and a sample size of 200. Its power is 51% at a reliability level of .75 and 80% for reliabilities of .95. However, power differences with the corrected means method and latent product method are only one to three percentage points across conditions. At reliabilities of .85 and a sample size of 200, the multi-group method (34% power due to discretization) and product indicators method (37% power due to standard error bias) have lower power.

Discussion

Study 1 raises concerns about the performance of the means, multi-group, and product indicators method, even at reliabilities of .85 that are conventional in the literature review (Table 1) and commonly considered good (Peterson 1994). In contrast, the corrected means, factor scores, and latent product method perform relatively well across conditions. Their parameter bias is under 10% and standard error bias below 5% (Feingold 2019; Muthén and Muthén 2002) at a sample size of 200 (and higher). There are also little differences in power and RMSE between these three methods. Main effect results offer similar conclusions (Web Appendix F).[2]

However, the estimated power to find a small-to-medium moderation effect of .20 (about the mean in the literature review, see Table 1) at a measure reliability level of .85 (about the mean) and a sample size of 200 (about the median) is only about 65% at best. To estimate required sample sizes for 80% power based on Study 1, we follow Schoemann et al. (2014) and extract fitted probabilities from a binary probit regression of the significance of the moderation effect (1 if it is statistically significant, 0 otherwise) on an intercept, the sample size, the dummy-coded reliability, the dummy-coded method and all interactions. The estimated required sample size is then the smallest sample for which the estimated likelihood (power) of a statistically significant moderation effect is at least 80%.

Table 6 reports the estimates. To find a moderation effect of .20 at a reliability of .85 and with 80% power, the corrected means, factor scores, and latent product methods need at

Table 6
Study 1: Required Sample Size Estimates To Estimate a .20 Moderation Effect With 80% Power

Reliability of Y, X and Z	3. Corrected means	4. Factor scores	5. Product indicators	6. Latent product
.75	449 [442, 455]	443 [436, 449]	537 [530, 544]	450 [443, 456]
.85	309 [305, 314]	309 [305, 314]	334 [329, 339]	312 [307, 317]
.95	215 [212, 218]	214 [211, 218]	217 [214, 220]	216 [213, 219]

Notes: Cells contain point estimates and 95% confidence intervals of the required minimum sample sizes to estimate a moderation effect of .20 (about the average in the literature review) with 80% power across methods and reliabilities of Y, X and Z. Estimates are based on a binary probit regression (Schoemann et al. 2014). The median sample size in the literature review is 202 and the mean reliability is .88 (Table 1).

least 312 observations. This requirement is more than 50% larger than the median sample size of 202 in the literature review and only met by 28% of studies in our literature review. Thus, larger sample sizes are needed to attain sufficient power. At a high reliability of .95, slightly more than 200 observations are sufficient for 80% power. Smaller reliabilities of .75 require even larger samples (e.g., ≥ 450 for latent product method). These results are in line with findings in the strategic management domain (Aguinis et al. 2017) and suggest that a substantive proportion of published moderation effects under investigation might be biased downward (due to the widespread use of the means method) and/or underpowered (due to moderation analysis in small samples).

STUDY 2A: ORDERED CATEGORICAL INDICATORS

Design

Study 2a extends Study 1 by using ordered categorical indicators rather than continuous indicators. The design is: 6 (Method) × 7 (Sample size) × 3 (Number of scale points of y, x and z: 7, 5 or 3). We follow Rhemtulla et al. (2012) and use thresholds based on Z-scores that equally divide ±2.5 standard deviations from the mean to transform the continuous indicators. We focus on seven-point scales (60% of the cases in the literature review), five-point scales (15%), and three-point scales (below 1%) explore boundary conditions. Overall, categorical indicators contain less information than continuous indicators do but Rhemtulla et al. (2012) find that indicators with five or more ordered categories perform similar to continuous indicators in non-moderation settings. Study 2a tests whether this holds for latent moderation.

Results

First, the bias of the multi-group and means methods increases when the number of scale points decreases. For instance, from 27% (7-point) to 38% (3-point) parameter bias for the means method (26% for continuous indicators in Study 1). Second, the factor scores method

and latent product method remain unbiased (parameter and standard error bias below 5%) across conditions and their RMSE and power levels are similar (e.g., RMSE of .37 for factor scores and .39 for the latent product method). However, power levels are lower than in Study 1. The latent product method has a power of 58%, 56% and 45% for seven-, five- and three-point scales at a reliability of .85 and sample size of 200, while it had 64% power in Study 1. Third, the corrected means and product indicators methods are biased for three-point scales (standard error bias up to 43% at a sample size of 200). However, and interestingly, the product indicators method has a standard error bias of 5% for at least five-point scales, whereas it had standard error bias of 10% at a sample size of 200 for continuous indicators (Study 1). In this simulation, categorical scales limit extreme values in the indicators, such as outliers, that are more likely to occur for continuous scales and become bigger issues due to indicator multiplication. In sum, although categorical indicators contain less information than continuous ones, leading to lower power, five-point and seven-point scales perform almost equally to continuous indicators in terms of unbiasedness for the factor scores method and latent product method. The factor scores method and latent product method outperform the corrected means method and product indicators method for three-point scales.

STUDY 2B: DISCRETE MODERATOR

Design

Study 2b extends Study 1 by focusing on a single discrete (binary) moderation indicator without measurement error (e.g., a country indicator or a manipulation dummy). This is the case for about a third of the moderation effects in the literature review (Web Appendix A). The design is: 6 (Method) × 7 (Sample size) × 3 (Reliability of Y, X and Z: .95, .85 or .75). Here, the multi-group method can use the moderator without discretization and we investigate how multi-group performs compared to the other methods in such a setting.

Results

First, the multi-group method is unbiased (bias under 2% across sample sizes) for a discrete moderator. Similarly, standard error biases are under 5% at a sample size of 200. Second, the bias of the means method reduces compared to Study 1 but persists (about 15% at a reliability of .85) unless reliabilities are .95 (bias about 5%). Third, the parameter and standard error biases of the product indicators method reduce compared to Study 1 (below 5% across reliabilities of .75 to .95 and for sample sizes of 200 and larger). These findings are due to fact that Study 2b only has measurement error in x and y, whereas Study 1 focused on measurement error in y, x and z. Fourth, corrected means, factor scores, product indicators, and latent product methods are unbiased (parameter and standard error bias up to 5%). RMSE (e.g., between .33 (factor scores and corrected means) and .35 (multi-group) at a reliability of .85 and sample size of 200) and power (70-71%) are similar under the investigated conditions for the unbiased methods. In sum, the multi-group method is a well-performing alternative to the corrected means, factor scores, product indicators, and latent product methods for binary moderators without measurement error.

STUDY 2C: CORRELATION OF X WITH Z

Design

Study 2c extends Study 1 by varying the correlation between X and Z (fixed to .20 in Study 1). Typically, X and Z are correlated in observational data and this might impact method performance (Grewal et al. 2004). The design is: 6 (Method) × 7 (Sample size) × 4 (Correlation of X with Z: 0, .20, .40, .60) with the correlation varying from 0 to .60, in line with the range in the literature review (Web Appendix A).

Results

First, the bias of the moderation effect for the multi-group method decreases when the correlation between X and Z increases, but the main effects (see Web Appendix I) become more biased (up to 86% at a correlation of .60). Second, increasing the correlation between X and Z from 0 to .60 decreases the moderation bias for the means method from 27% to 21%. This is due to the higher reliability of product terms for correlated components (see Equation (6)). Third, the corrected means, factor scores, and latent product method are unbiased across conditions (parameter bias below 3% and standard error bias below 5%) whereas the product indicators method has standard error bias of 9-10%. The unbiased methods have similar RMSE and power levels (e.g., RMSE between .40 (factor scores) and .42 (latent product method) at a correlation of .60 and sample size of 200). Fourth, the power of the moderation effect increases when the correlation between X and Z increases from 0 to .60, from 62% to 79% for the latent product method. Thus, the increase in power of the moderation effect due to the increased reliability of the product term trades off against the decrease in power due to multicollinearity. However, consistent with Grewal et al. (2004), the power of the main effects decreases due to multicollinearity (from 67% to 47% for the main effects; see Web Appendix I). In sum, the corrected means, factor scores, and latent product method are unbiased under the investigated conditions. Higher correlation between X and Z increases power to find a moderation effect but decreases power of the main effects.

STUDY 3: UNEQUAL INDICATOR LOADINGS

Design

Study 3 extends Study 1 by focusing on indicators of the latent variables that differ in their loadings. Because the multi-group method and the means method are biased across conditions in Studies 1, 2a and 2c, upon which the following studies build, Studies 3 and 4a-c focus on the comparison between the remaining methods: factor scores, corrected means,

product indicators, and latent product. The design is: 4 (Method) × 7 (Sample size) with unequal indicators in all cells: λ_{x1}= 1, λ_{x2} = 1.5, λ_{x3} = .50 (and analogous for Z and Y). We hold indicator measurement error variances constant such that measure (composite) reliabilities are equal with those from Study 1 and to make sure that differences between equal and unequal loading conditions are not confounded with differences in measure reliability. We expect the factor scores method and latent product method to perform best because they freely estimate loadings. The corrected means method assumes that all indicators are equally good representatives of their underlying latent factors, and Cronbach's alpha underestimates measure reliability if this assumption is violated (McNeish and Wolf 2020). This might lead to measurement error corrections that bias the estimates upward.

Results

First, as expected, the corrected means method has biased moderation and main effect estimates, at least 20% even at large sample sizes of 1,500. Second, the factor scores and latent product methods that freely estimate indicator loadings perform best, with parameter and standard error biases under 5% at sample sizes of 200 (and higher). Their RMSE and power are similar (e.g., 69% power of the factor scores method and 68% power of the latent product method at a sample size of 200). Third, the product indicators method has a low parameter bias as the factor scores method and the latent product method have, but a higher standard error bias (e.g., 12% at a sample size of 200). In sum, the factor scores method and latent product method perform best for unequal indicator loadings.

STUDY 4A: NON-NORMALLY DISTRIBUTED INDICATORS

Design

Studies 4a-c investigate situations where model assumptions of all focal methods are violated, unlike Study 3 that only violates assumptions of the corrected means method. Study

4a focuses on non-normality distributed x and z, which is common when measuring constructs such as customer satisfaction (Peterson and Wilson 1992). The design is: 4 (Method) × 7 (Sample size) × 2 (Skewness / excess kurtosis of X and Z: 1 / 2 or 3 / 10). Skewness and excess kurtosis are conventional metrics of non-normality. Both are zero for normally distributed variables (Oliveira et al. 2016). Because we could not determine skewness and excess kurtosis in our literature review, we use about the 75^{th} and 95^{th} percentiles from a recent existing review in psychology (Cain et al. 2017, p. 1720). The procedure described in Vale and Maurelli (1983) generates non-normal latent variables X and Z that reflect in non-normal indicators. Previous research concluded that non-zero skewness and excess kurtosis in variables lead to overestimated zero-order correlations (Bishara and Hittner 2015) but underestimated standard errors (Finch et al. 1997). Yet, there might be differences between methods. Biased reliability estimates due to non-normality can bias the corrected means method (Sheng and Sheng 2012). The product indicators method was found to be robust for different latent variable distributions (Marsh et al. 2004) although taking multiple indicator products might also exacerbate bias due to non-normality. The latent product method does not use (algebraic) multiplications of indicators so it might perform better, but severe non-normality can still hamper the ability of the mixture distribution to approximate the indicator distribution (Klein and Moosbrugger 2000).

Results

First, all methods are biased (up to 19% for the corrected means method) in presence of severe non-normality in x and z (i.e., skewness of X and Z is 3 and excess kurtosis is 10). One exception is the product indicators method with 5% parameter bias at a sample size of 200. Second, standard errors of all methods are also biased, including those of the product indicators method (standard error bias of 32%). Third, for moderately non-normally distributed indicators (i.e., skewness of X and Z is 1 and excess kurtosis is 2), the factor

scores method and latent product method have biases under 5%. RMSE and power levels are similar (e.g., RMSE .35-.38 at a sample size of 200). In sum, the expectation of severe non-normally distributed indicators with skewness and excess kurtosis might call for the product indicators method even though its statistical conclusion validity might be questionable due to biased standard errors.

STUDY 4B: CORRELATED MEASUREMENT ERRORS

Design

Study 4b focuses on another type of misspecification: correlated measurement errors. The design is: 4 (Method) × 7 (Sample size) × 3 (Measurement error correlation: x with y, x with z or x with x). Correlated measurement errors can occur due to omitted variables in the measurement model such as method factors or response tendencies (Baumgartner and Weijters 2017). We focus on three types of measurement error correlations. First, we generate error correlations between indicators of x and y (denoted 'x with y'). Evans (1985) and Siemsen et al. (2010) showed in the context of the means method that measurement error correlations between x and y do not bias moderation effects upward but can bias them downward depending on the magnitude of measurement error correlation. However, it is unclear whether these results hold for the main effects, the other methods, and for other measurement error correlations. Henceforth, the design also includes measurement error correlation between moderation indicators x and z and for indicators of X with other indicators of X (denoted 'x with x'). For brevity, we do not focus on measurement error correlations of z with y (analogous to x with y) and z with z (analogous to x with x). The measurement error correlation in all cells is .50. To generate the correlated measurement errors for x with y (analogous for x with z), we correlate indicator x1 with y1, x2 with y2, and x3 with y3. Measurement error correlations of x with x intercorrelate all three indicators of X.

Results

First, measurement error correlations of .50 between x and y bias the main effect estimate of X up to 50% for all methods (Web Appendix L has details) even though the moderation effect is unbiased (under 5%). This extends what was previously found for the means method (Evans 1985; Siemsen et al. 2010). Second, under the investigated conditions, measurement error correlations of x with z yield parameter biases under 10% for the moderation and main effects across methods, much less than for measurement error correlations between x and y. However, the standard error bias of the product indicators method is 9% whereas the standard error bias of the corrected means, factor scores and latent product methods is 2-4%. Third, measurement error correlations of x with x also severely bias the moderation and main effects of the corrected means method, product indicators method and latent product method for about 21%. However, the bias is 12%, about 9% less, for the factor scores method. One reason for this result might be that the two-step estimation of the factor scores method, compared to one-step or simultaneous estimation of the latent product method, is more robust to misspecification in the measurement model (Devlieger and Rosseel 2017; Rosseel 2020; Smid and Rosseel 2020). Thus, under the investigated conditions, correlated measurement error biases all methods. The bias is most severe for measurement error correlations of predictors with outcomes (e.g., x with y).

STUDY 4C: STRUCTURAL MODEL IS MISSPECIFIED

Design

Study 4c focuses on misspecification of the structural model. The design is: 4 (Method) × 7 (Sample size) × 4 (Correlation of X with Z: 0, .20, .40, .60). It generates the data with a U-shape of X (i.e., $Y = \beta_1 X + \beta_2 Z + \beta_4 X^2$) and uses the structural model in Equation (1) for estimation. Because moderation product terms and squared terms are generally correlated due to their common lower order components if they are not manipulated (Ganzach 1997), the

design varies the correlation between X and Z. Although we expect little differences between the methods, it is difficult to quantify bias and resulting type I error analytically.

Results

First, when X and Z are uncorrelated, we find that the methods yield unbiased (≤ 2%) moderation effects. Bias for all methods is just under 10% when X and Z are correlated .20. Second, when the correlation between X and Z increases, the bias due to misspecification increases, for instance to 19% for the latent product method and at a correlation of X with Z of .60 and a sample size of 200. Third, standard errors of all methods are biased between 5% and 15% across conditions, even at large sample sizes of 1,500. Fourth, all methods have type I error ≥ 5% across conditions, about 20% at a correlation of .20 and a sample size of 200, which further increases if the correlation between X and Z or sample size increases.

GENERAL DISCUSSION

We compared six methods for latent moderation analysis and provide several recommendations for latent moderation analysis. First, the choice between five out of the six methods is at the researcher's discretion when reliabilities of moderation variables approach one. Although the multi-group method is biased for over 20% when the indicators of the moderator are continuous, the parameter bias of the corrected means, factor scores, product indicators, and latent product methods across sample sizes is under 2% and the standard error bias under 5% when the reliability of Y, X and Z was a high .95 (Study 1). The parameter bias of the means method is then 8% (and standard error bias 3%), which might be acceptable (Table 2). Even more, RMSE and power differences were small. The closer the reliabilities of the moderating variables are to one, the more similar the performance of five out of the six methods becomes.

Yet, reliabilities of moderation variables approaching one is rare in practice: the mean reliability in the literature review was .88 (Table 1) and only 13% of moderation tests had reliabilities of the moderation variables ≥ .95. Thus, our findings and recommendation are in contrast with the 94% use of the means method in the literature review (Table 1). It is well known that ignoring measurement error can bias parameter estimates (Grewal et al. 2004; Spearman 1904; Wooldridge 2015). Study 1 shows the bias of the means method once more and our Monte Carlo studies quantify it in the latent moderation context: the moderation effect bias of the means method is 40% and 25% respectively at reliabilities of .75 and .85.

Second, the factor scores method and latent product method are recommended across most investigated conditions (Table 5). When indicators are continuous (Study 1) or seven-, five- or three-point ordered categorical (Study 2a), or when the moderator is binary (Study 2b), and across reliabilities between .75 and .95 (Studies 1 and 2b), the factor scores method and latent product method have parameter and standard error biases ≤ 5%. The bias remains small when the correlation between moderation variables increases from 0 to .60 (Study 2c) and for unequal indicator loadings (Study 3). We conclude from the small RMSE and power differences within the conditions of our studies that the choice between the factor scores method and the latent product method is mostly at the researcher's discretion. Method accessibility can then be relevant. Factor scores are available in most general statistical software packages. The latent product method is to our knowledge currently only available in Mplus (Muthén and Muthén 2019) and in an R package (Umbach et al. 2017). A follow-up study in Web Appendix N compares both latent product implementations and recommends Mplus in terms of performance, computation time, and the range of possible models that can be estimated. One key researcher decision to use the latent product method is the number of mixture components. A follow-up study in Web Appendix O shows that the default in Mplus is adequate to estimate a single moderation effect (Klein and Moosbrugger 2000).

When using the factor scores method, decisions need to be made about the type of measurement model and the factor scores estimation method. We draw from Skrondal and Laake (2001) and Devlieger et al. (2016) and our own analyses to recommend the following "two-step factor scores (TSFS)" method for latent moderation analysis.

Step 1: conduct a confirmatory factor analysis with the outcome (Y) as a single factor (1-CFA) and extract Bartlett factor scores. The 1-CFA uses optimal indicator weighting and Bartlett factor scores account for measurement error in the outcome. Then conduct a separate confirmatory factor analysis for the predictors (X and Z) simultaneously with two factors that are allowed to correlate (2-CFA, no cross-loadings) and extract regression factor scores to assure optimal indicator weighting and account for measurement error in predictors.

Step 2: compute the product term from the factor scores of the predictors (multiply) and estimate moderation and main effects with the target regression or path model. We demonstrate that this TSFS method for latent moderation analysis performs well across the examined range of conditions, and about as well as the latent product method that estimates the measurement and structural models simultaneously. Web Appendix P contains a follow-up simulation study that empirically examines the harm of using different factor scores methods than those recommended here.

Third, the multi-group, corrected means, and product indicators method are best reserved for specific settings. The multi-group method can be used for discrete moderators with bias less than 5% although the corrected means, factor scores, product indicators, and latent product method perform similarly. The corrected means method was found by others to perform well and similarly to the latent product method for single-indicators (Hsiao et al. 2021). In that case, the factor scores method cannot be used. Still, if a single-indicator that contains measurement error is available it becomes more difficult to estimate unreliability and hence to account for it, compared to multi-indicator measures for which reliability

estimators are readily available (Kamakura 2015). We refer to Pieters (2017, pp. 699-700) and the references therein for guidance. We identify one setting in which the product indicators method outperforms the factor scores method and the latent product method. The product indicators method had an estimated parameter bias of about 5% (parameter bias of 14% for the factor scores method and the latent product method) when the moderation variables had a skewness of 3 and excess kurtosis of 10 and at a sample size of 200 (Study 4a). Yet, standard errors of the product indicators method, as well as those of the other methods remain biased (e.g., 32% standard error bias for the latent product method at a sample size of 200), which can harm statistical conclusion validity. Overall, these recommendations should provide actionable guidelines for method use. Web Appendix B overviews sample code for method implementation in SPSS, Stata, R and Mplus, made available on OSF: https://osf.io/py7jx/?view_only=5d921a6658cf402a80bd1d4996665331.

There are situations when the corrected means, factor scores, product indicators, and latent product methods all perform poorly. First, although we showed that correlations between (latent variables) X and Z up to .60 have a negligible effect on the bias of these methods (Study 2c), they can be biased when measurement errors of individual indicators (e.g., x and z) are correlated, independent of the correlation between X and Z. This may occur, for instance, when indicators of X and/or Z are regular and reversed items. Then, the measurement model needs to be adapted (e.g., Baumgartner and Weijters 2017; Weijters et al. 2013), such as by introducing a method factor or having specific errors correlate, before applying the methods that we have compared here. Second, if the true effect of X on Y is U-shaped (polynomial) but not specified (Hutchinson et al. 2000), not only the means method (Ganzach 1997), but all methods perform poorly (Study 4c). That is, if the data generating process is a U-shape effect of X on Y, using a specification of the moderation without the polynomial X^2 leads to type I errors for all methods. Then, a non-zero moderation effect

between X and Z might be observed, whereas none exists in the data, which can avoided by examining moderation and curvilinear effects simultaneously (Ganzach 1997).

Among our findings, the small differences in performance between the TSFS method and the latent product method across the focal conditions are noteworthy. One might have expected that joint estimation of the measurement and structural models by the latent product method should empirically perform better than the TSFS method. Recent research in the non-moderation context has been drawing attention to the role of two-step vs. conventional simultaneous estimation of latent variable models (Devlieger and Rosseel 2017; Rosseel 2020; Smid and Rosseel 2020).[3] Conceptually, the TSFS method matches the estimation of the measurement and structural models as a combination of two separate models (Anderson and Gerbing 1988). Empirically, one advantage of two-step estimation is that measurement model misspecification might lead to less structural model bias, or vice versa (Devlieger and Rosseel 2017). Our Study 2c found this to be the case in the context of within-construct correlated measurement errors, although the reduction in bias of the factor scores method compared to the latent product method was a modest 9% under the investigated conditions. Moreover, two-step methods might have less convergence issues or ineligible solutions than simultaneous estimation methods do (Smid and Rosseel 2020). Follow-up analyses of our Study 1 find that although non-convergence was rare, all replications converged for the TSFS method and the corrected means method (both two-step methods). In contrast, 2.4% of replications for the product indicators method and less than 1% of replications for the latent product method (both one-step methods) did not converge. Among non-converging replications, small sample sizes of 100 or 150 (about 84%) and low reliabilities of .75 (about 86% of non-converging replications) were most common, which might support the use of two-step estimation to avoid convergence issues of simultaneous estimation in such settings

(Rosseel 2020). In sum, the TSFS method for latent moderation analysis is accessible and its estimates have a small bias with low variance across a large range of conditions.

With this foundation, our study opens several avenues for further research. First, one might investigate the performance of Bayesian estimation, which might do well in small samples and facilitates the incorporation of prior information, potentially resulting in more precise estimates and moderation tests with higher power. Asparouhov and Muthén (2021) conduct simulation studies for the latent product method. Second, although this research studied both random and correlated indicator measurement error, it does not focus on methods to account for correlated measurement errors. If correlated measurement errors are expected, the latent product method might be preferred over the factor scores method because it uses separate factor analyses for predictors and outcomes in which error correlations between predictors and outcomes cannot be accounted for. The latent product method estimates the measurement and structural models simultaneously. Further research might investigate this. We refer to Baumgartner and Weijters (2017) for an overview of models to account for correlated measurement errors in a non-moderation setting. Third, although the Monte Carlo simulations study a variety of conditions, including settings that violate assumptions, the simulations can be extended further. For instance, the question remains how the methods perform for multi-level or multi-time data and fixed or random effects models. Similarly, method performance in case of (latent) instrumental variables can be assessed.

In sum, it is hard to justify the continued use of the means method for latent moderation analysis unless measurement reliabilities approach one. Researchers are well advised to apply other methods for latent moderation analysis such as the two-step factor scores (TSFS) method and the latent product method. We hope that our recommendations improve moderation theory testing and help marketing researchers planning their next latent moderation studies.

REFERENCES

Aguinis, Herman, Jeffrey R. Edwards, and Kyle J. Bradley (2017), "Improving Our Understanding of Moderation and Mediation in Strategic Management Research," *Organizational Research Methods*, 20 (4), 665-85.

Anderson, James C. and David W. Gerbing (1988), "Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach," *Psychological Bulletin*, 103 (3), 411-23.

Asparouhov, Tihomir and Bengt O. Muthén (2021), "Bayesian Estimation of Single and Multilevel Models with Latent Variable Interactions," *Structural Equation Modeling: A Multidisciplinary Journal*, 28 (2), 314-28.

Atasoy, Ozgun and Carey K. Morewedge (2017), "Digital Goods Are Valued Less Than Physical Goods," *Journal of Consumer Research*, 44 (6), 1343-57.

Auh, Seigyoung, Bulent Menguc, Constantine S. Katsikeas, and Yeon Sung Jung (2019), "When Does Customer Participation Matter? An Empirical Investigation of the Role of Customer Empowerment in the Customer Participation–Performance Link," *Journal of Marketing Research*, 56 (6), 1012-33.

Baumgartner, Hans and Bert Weijters (2017), "Measurement Models for Marketing Constructs," in *Handbook of Marketing Decision Models*, Berend Wierenga and Ralf van der Lans, eds. 2nd ed. Cham, Switzerland: Springer.

Bishara, Anthony J. and James B. Hittner (2015), "Reducing Bias and Error in the Correlation Coefficient Due to Nonnormality," *Educational and Psychological Measurement*, 75 (5), 785-804.

Bollen, Kenneth A. (1989), Structural Equations with Latent Variables. New York: Wiley.

Busemeyer, Jerome R. and Lawrence E. Jones (1983), "Analysis of Multiplicative Combination Rules When the Causal Variables Are Measured with Error," *Psychological Bulletin*, 93 (3), 549-62.

Cain, Meghan K., Zhiyong Zhang, and Ke-Hai Yuan (2017), "Univariate and Multivariate Skewness and Kurtosis for Measuring Nonnormality: Prevalence, Influence and Estimation," *Behavior Research Methods*, 49 (5), 1716-35.

Charles, Eric P. (2005), "The Correction for Attenuation Due to Measurement Error: Clarifying Concepts and Creating Confidence Sets," *Psychological Methods*, 10 (2), 206-26.

Cohen, Jacob (1988), Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, Jacob, Patricia Cohen, Stephen G. West, and Leona S. Aiken (2003), Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), "Maximum Likelihood from Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 39 (1), 1-38.

Devlieger, Ines, Axel Mayer, and Yves Rosseel (2016), "Hypothesis Testing Using Factor Score Regression: A Comparison of Four Methods," *Educational and Psychological Measurement*, 76 (5), 741-70.

Devlieger, Ines and Yves Rosseel (2017), "Factor Score Path Analysis," *Methodology*, 13 (1), 31-38.

Dimitruk, Polina, Karin Schermelleh-Engel, Augustin Kelava, and Helfried Moosbrugger (2007), "Challenges in Nonlinear Structural Equation Modeling," *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 3 (3), 100-14.

Eisend, Martin (2015), "Have We Progressed Marketing Knowledge? A Meta-Meta-Analysis of Effect Sizes in Marketing Research," *Journal of Marketing*, 79 (3), 23-40.

Evans, Martin G. (1985), "A Monte Carlo Study of the Effects of Correlated Method Variance in Moderated Multiple Regression Analysis," *Organizational Behavior and Human Decision Processes*, 36 (3), 305-23.

Feingold, Alan (2019), "Time-Varying Effect Sizes for Quadratic Growth Models in Multilevel and Latent Growth Modeling," *Structural Equation Modeling: A Multidisciplinary Journal*, 26 (3), 418-29.

Finch, John F., Stephen G. West, and David P. MacKinnon (1997), "Effects of Sample Size and Nonnormality on the Estimation of Mediated Effects in Latent Variable Models," *Structural Equation Modeling: A Multidisciplinary Journal*, 4 (2), 87-107.

Foldnes, Njål and Knut Arne Hagtvet (2014), "The Choice of Product Indicators in Latent Variable Interaction Models: Post Hoc Analyses," *Psychological Methods*, 19 (3), 444-57.

Ganzach, Yoav (1997), "Misleading Interaction and Curvilinear Terms," *Psychological Methods*, 2 (3), 235-47.

Germann, Frank, Peter Ebbes, and Rajdeep Grewal (2015), "The Chief Marketing Officer Matters!," *Journal of Marketing*, 79 (3), 1-22.

Grewal, Rajdeep, Joseph A. Cote, and Hans Baumgartner (2004), "Multicollinearity and Measurement Error in Structural Equation Models: Implications for Theory Testing," *Marketing Science*, 23 (4), 519-29.

Hallquist, Michael N. and Joshua F. Wiley (2018), "MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus," *Structural Equation Modeling: A Multidisciplinary Journal*, 25 (4), 621-38.

Hsiao, Yu-Yu, Oi-Man Kwok, and Mark H. C. Lai (2021), "Modeling Measurement Errors of the Exogenous Composites from Congeneric Measures in Interaction Models," *Structural Equation Modeling: A Multidisciplinary Journal*, 28 (2), 250-60.

Hunter, John E. and Frank L. Schmidt (2004), Methods of Meta-Analysis: Correcting Error and Bias in Research Findings (2nd ed.). Thousand Oaks, CA: Sage.

Hutchinson, J. Wesley, Wagner A. Kamakura, and John G. Lynch (2000), "Unobserved Heterogeneity as an Alternative Explanation for "Reversal" Effects in Behavioral Research," *Journal of Consumer Research*, 27 (3), 324-44.

Irwin, Julie R. and Gary H. McClelland (2001), "Misleading Heuristics and Moderated Multiple Regression Models," *Journal of Marketing Research* 38 (1), 100-09.

---- (2003), "Negative Consequences of Dichotomizing Continuous Predictor Variables," *Journal of Marketing Research*, 40 (3), 366-71.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2013), An Introduction to Statistical Learning (1st ed.). New York: Springer.

Jöreskog, Karl G. and Fan Yang (1996), "Nonlinear Structural Equation Models: The Kenny-Judd Model with Interaction Effects," in *Advanced Structural Equation Modeling: Issues and Techniques*, G. Marcoulides and R. Schumacker, eds. New York, NY: Psychology Press.

Kamakura, Wagner A. (2015), "Measure Twice and Cut Once: The Carpenter's Rule Still Applies," *Marketing Letters*, 26 (3), 237-43.

Kenny, David A. and Charles M. Judd (1984), "Estimating the Nonlinear and Interactive Effects of Latent Variables," *Psychological Bulletin*, 96 (1), 201-10.

Klein, Andreas and Helfried Moosbrugger (2000), "Maximum Likelihood Estimation of Latent Interaction Effects with the LMS Method," *Psychometrika*, 65 (4), 457-74.

Lastovicka, John L. and Kanchana Thamodaran (1991), "Common Factor Score Estimates in Multiple Regression Problems," *Journal of Marketing Research*, 28 (1), 105-12.

Lin, Guan-Chyun, Zhonglin Wen, Herbert W. Marsh, and Huey-Shyan Lin (2010), "Structural Equation Models of Latent Interactions: Clarification of Orthogonalizing and Double-Mean-Centering Strategies," *Structural Equation Modeling: A Multidisciplinary Journal*, 17 (3), 374-91.

Marsh, Herbert W., Zhonglin Wen, and Kit-Tai Hau (2004), "Structural Equation Models of Latent Interactions: Evaluation of Alternative Estimation Strategies and Indicator Construction," *Psychological Methods*, 9 (3), 275-300.

McClelland, Gary H., Julie R. Irwin, David Disatnik, and Liron Sivan (2017), "Multicollinearity Is a Red Herring in the Search for Moderator Variables: A Guide to Interpreting Moderated Multiple Regression Models and a Critique of Iacobucci, Schneider, Popovich, and Bakamitsos (2016)," *Behavior Research Methods*, 49 (1), 394-402.

McNeish, Daniel and Melissa Gordon Wolf (2020), "Thinking Twice About Sum Scores," *Behavior Research Methods*, 52 (6), 2287-305.

Moosbrugger, Helfried, Karin Schermelleh-Engel, and Andreas Klein (1997), "Methodological Problems of Estimating Latent Interaction Effects," *Methods of Psychological Research Online*, 2 (2), 95-111.

Muthén, Linda K. and Bengt O. Muthén (2002), "How to Use a Monte Carlo Study to Decide on Sample Size and Determine Power," *Structural Equation Modeling: A Multidisciplinary Journal*, 9 (4), 599-620.

---- (2019), Mplus User's Guide (8th ed.). Los Angeles, CA: Muthén & Muthén.

Oliveira, Amílcar, Teresa Oliveira, and Antonio Seijas-Macías (2016), "Evaluation of Kurtosis into the Product of Two Normally Distributed Variables," *AIP Conference Proceedings*, 1738 (1), 1-4.

Peterson, Robert A. (1994), "A Meta-Analysis of Cronbach's Coefficient Alpha," *Journal of Consumer Research*, 21 (2), 381-91.

Peterson, Robert A. and William R. Wilson (1992), "Measuring Customer Satisfaction: Fact and Artifact," *Journal of the Academy of Marketing Science*, 20 (1), 61.

Pieters, Rik (2017), "Meaningful Mediation Analysis: Plausible Causal Inference and Informative Communication," *Journal of Consumer Research*, 44 (3), 692-716.

R Core Team (2020), "R: A Language and Environment for Statistical Computing." Vienna, Austria: R Foundation for Statistical Computing.

Rhemtulla, Mijke, Patricia É Brosseau-Liard, and Victoria Savalei (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), 354-73.

Rosseel, Yves (2012), "lavaan: An R Package for Structural Equation Modeling," *Journal of Statistical Software*, 48 (2), 1-36.

---- (2020), "Small Sample Solutions for Structural Equation Modeling," in *Small Sample Solutions: A Guide for Applied Researchers and Practitioners*, Rens Van de Schoot and Milica Miočević, eds. New York: Routhledge.

Schoemann, Alexander M., Patrick Miller, Sunthud Pornprasertmanit, and Wei Wu (2014), "Using Monte Carlo Simulations to Determine Power and Sample Size for Planned Missing Designs," *International Journal of Behavioral Development*, 38 (5), 471-79.

Sheng, Yanyan and Zhaohui Sheng (2012), "Is Coefficient Alpha Robust to Non-Normal Data?," *Frontiers in Psychology*, 3 (34), 1-13.

Siemsen, Enno, Aleda Roth, and Pedro Oliveira (2010), "Common Method Bias in Regression Models with Linear, Quadratic, and Interaction Effects," *Organizational Research Methods*, 13 (3), 456-76.

Singh, Sunil K., Detelina Marinova, and Jagdip Singh (2020), "Business-to-Business E-Negotiations and Influence Tactics," *Journal of Marketing*, 84 (2), 47-68.

Skrondal, Anders (2000), "Design and Analysis of Monte Carlo Experiments: Attacking the Conventional Wisdom," *Multivariate Behavioral Research*, 35 (2), 137-67.

Skrondal, Anders and Petter Laake (2001), "Regression among Factor Scores," *Psychometrika*, 66 (4), 563-75.

Smid, Sanne C. and Yves Rosseel (2020), "SEM with Small Samples: Two-Step Modeling and Factor Score Regression Versus Bayesian Estimation with Informative Priors," in *Small Sample Solutions: A Guide for Applied Researchers and Practitioners*, Rens Van de Schoot and Milica Miočević, eds. New York: Routhledge.

Spearman, C. (1904), "The Proof and Measurement of Association between Two Things," *The American Journal of Psychology*, 15 (1), 72-101.

Umbach, Nora, Katharina Naumann, Holger Brandt, and Augustin Kelava (2017), "Fitting Nonlinear Structural Equation Models in R with Package nlsem," *Journal of Statistical Software*, 77 (1), 1-20.

Vale, C. David and Vincent A. Maurelli (1983), "Simulating Multivariate Nonnormal Distributions," *Psychometrika*, 48 (3), 465-71.

Van Smeden, Maarten, Timothy L. Lash, and Rolf H. H. Groenwold (2019), "Reflection on Modern Methods: Five Myths About Measurement Error in Epidemiological Research," *International Journal of Epidemiology*, 49 (1), 338-47.

Weijters, Bert, Hans Baumgartner, and Niels Schillewaert (2013), "Reversed Item Bias: An Integrative Model," *Psychological Methods*, 18 (3), 320-34.

Wooldridge, Jeffrey M. (2015), Introductory Econometrics: A Modern Approach (6th ed.). Boston, MA: Cengage Learning.

Yuan, Ke-Hai, Ying Cheng, and Wei Zhang (2010), "Determinants of Standard Errors of MLEs in Confirmatory Factor Analysis," *Psychometrika*, 75 (4), 633-48.

FOOTNOTES

[1] A reliability estimator of X^2 is the square of the reliability of X (Dimitruk et al. 2007), so by definition lower than the reliability of X and usually lower than the reliability of X and Z (if reliability Z = reliability X) unless X and Z are uncorrelated.

[2] There is a possibility that the performance of the methods differs for specific subsets of the data. For instance, methods that are more heavily parameterized (like the product indicators method and the latent product method) might be more prone to fitting idiosyncrasies in the data (e.g., sampling error) instead of recovering the true moderation effect, which is undesirable. We perform ten-fold cross-validation (James et al. 2013, p. 181; Singh et al. 2020) to examine this. We use the focal four performance criteria to compare the methods. Preferred methods should only have small differences in terms of the in-sample performance criteria and those based on ten-fold cross validation. Web Appendix F summarizes cross-validation results of Study 1 that have only small differences with the in-sample performance. This is encouraging and rules out overfitting. Because we find little substantive differences between performance of the methods in-sample and based on ten-fold cross-validation, Studies 2a-c, 3 and 4a-c do not perform cross-validation.

[3] A Web of Science citation analysis of Anderson and Gerbing (1988), an early study advocating for two-step estimation of measurement and structural models, showed that the total number of citations per year increased from 149 in 2000 to 244 in 2005, 614 in 2010, 1,140 in 2015 and 2,143 in 2020. This might signal a more general trend of using two-step estimation for latent variable models.

SIX METHODS FOR LATENT MODERATION ANALYSIS IN MARKETING RESEARCH: A COMPARISON AND GUIDELINES

WEB APPENDIX

Note: The Web Appendix contains background material. To limit its length, an OSF repository at https://osf.io/py7jx/?view_only=5d921a6658cf402a80bd1d4996665331 contains additional details to which this Web Appendix refers.

Web Appendix A: Literature review

The purpose of the literature review is twofold. First, it seeks to assess the usage of the methods for moderation analysis in marketing research. Second, the results serve as an input for realistic Monte Carlo simulation studies that assess the performance of the methods.

We searched all 1,144 articles published in five volumes (2015-2019) of the premier marketing outlets *Journal of Marketing Research (JMR)*, *Journal of Marketing (JM)*, *Journal of Consumer Research (JCR)* and *Marketing Science (Mark. Sci.)* for keywords related to moderation. Specifically, the search was "moderat OR interact OR U-shape." The objective was to select articles with at least one moderation effect that could in principle be estimated with each of the methods. Hence, we manually selected articles with moderation effects (including quadratic effects or U-shapes) that had reliability information available for at least one of the interacting variables. Ultimately, we identified 656 moderation effects in 293 studies in 164 articles, for an average of 4 (Med = 2; Mo = 1; SD = 4.41; range = 1-30) effects per article. This procedure selected about 13% of all published 2015-2019 articles in *JMR*, 19% of *JM*, 24% of *JCR*, and 1% of *Mark. Sci.*, an overall 14% across focal outlets and years. Thus, moderation effects in face of measurement error are widespread in contemporary marketing research. Table WA1 has a detailed breakdown across outlets and volumes. The OSF repository has a full list of the 164 focal articles.

How widely are the six methods for moderation analysis used? Table WA2 shows that 154 (94%) out of the 164 articles used means. Thus, the vast majority of moderation effect estimations did not fully account for measurement error, despite unreliability information being available for at least one of the interacting variables. Four articles used multi-group, of which all used naturally categorical variables as grouping variable, such as a manipulation in Study 2 of Reinholtz et al. (2015), or followed the multi-group analysis up with an analysis of continuous interaction. Thus, moderation articles followed the advice of Irwin and

McClelland (2001, 2003) to avoid the discretization of continuous variables, which is good. One article (Katsikeas et al. 2018) used (Study 1) corrected means to estimate an interaction between exploitative and explorative learning (both 3 indicators, reliability = .84) and decision making complexity on performance (7 formative indicators) among 378 salespeople. Only seven articles used factor scores. For instance, Wathne et al. (2018) investigated how the effect of supplier incremental investments (3 indicators, reliability = .90) on ex-post transaction costs (7 indicators, reliability = .94) was moderated by reseller selection efforts (6 indicators, reliability = .93) in a sample of 100 resellers and supplier pairs in the building materials industry. This study used factor scores for all the multi-indicator scales (p. 709). Fürst et al. (2017) studied the effect of multichannel task differentiation (4 indicators, reliability = .82) on multichannel horizontal conflict (3 indicators, reliability = .88) among 329 key informants, and how it is moderated by customer cross-channel buying (4 indicators, reliability = .86). It is the only article in our sample that used product indicators, and it used four pairs as recommended by Marsh et al. (2004). Auh et al. (2019) studied several moderators of the relationship between customer participation (5 indicators, reliability = .89) and satisfaction (4 indicators, reliability = .93) effect. For instance, a latent product analysis (Klein and Moosbrugger 2000) among 891 customer-banker pairs found support for a negative moderation of customer orientation (5 indicators, reliability = .88). In sum, moderation tests that fully account for measurement error to estimate the moderation effect are rare in our sample.

The median study sample size is 202, quite close to the mean of 183 in a recent review of mediation analyses in volumes 41 and 42 (2014-2016) of *JCR* (Pieters 2017). The 25% and 75% percentiles of the sample size are respectively 129 and 329.

Table WA1
Literature Review: Article Outlet and Publication Year

Outlet	Publication year (volume & issues)	Total # of articles	# of focal articles	% focal out of total
JMR	2019 (56-1 to 56-6)	60	11	18%
	2018 (55-1 to 55-6)	60	11	18%
	2017 (54-1 to 54-6)	64	6	9%
	2016 (53-1 to 53-6)	67	7	10%
	2015 (52-1 to 52-6)	58	4	7%
	Total	309	39	13%
JM	2019 (83-1 to 83-6)	48	12	25%
	2018 (82-1 to 82-6)	51	9	18%
	2017 (81-1 to 81-6)	48	8	17%
	2016 (80-1 to 80-6)	42	9	21%
	2015 (79-1 to 79-6)	37	5	14%
	Total	226	43	19%
JCR	2019 (45-5 to 46-4)	68	19	28%
	2018 (44-5 to 45-4)	70	18	26%
	2017 (43-5 to 44-4)	79	11	14%
	2016 (42-5 to 43-4)	60	14	23%
	2015 (41-5 to 42-4)	58	18	31%
	Total	335	80	24%
Mark. Sci.	2019 (38-1 to 38-6)	52	1	2%
	2018 (37-1 to 37-6)	53	0	0%
	2017 (36-1 to 36-6)	55	0	0%
	2016 (35-1 to 35-6)	55	0	0%
	2015 (34-1 to 34-6)	59	1	2%
	Total	274	2	1%
Total		1,144	164	14%

Notes: *JMR* is the *Journal of Marketing Research*, *JM* is the *Journal of Marketing*, *JCR* is the *Journal of Consumer Reseach* and *Mark. Sci.* is *Marketing Science.* Total # of articles is based on a *Web of Science* search with query "SO = ('Journal of Marketing' OR 'Journal of Marketing Research' OR 'Journal of Consumer Research' OR 'Marketing Science') AND PY = 2015-2019". Percentages across outlets are 13%, 13%, 10%, 16% and 19% for 2015-2019 respectively.

Table WA2
Literature Review of 656 Moderation Analyses in Four Marketing Outlets 2015-2019

Category	Result
Articles	
Number of articles	164
Multi-group method	4 (2%)
Means method	154 (94%)
Corrected means method	1 (1%)
Factor scores method	7 (4%)
Product indicators method	1 (1%)
Latent product method	1 (1%)
Studies	
Number of studies	293
Sample size	M = 5,581; Mdn = 202 (SD = 57,493; range = 37-951,819)
Moderation effects	
Number of moderation effects	656
Test of a moderation hypothesis	637 (97%)
Test of a U-shape hypothesis	19 (3%)
Measured X and Z	266 (41%)
Manipulated X or Z	390 (59%)
Effect sizes & correlations	
Effect size of the main effects (344 out of 1,312 effects)	M = .20; M_w = .21; Mdn = .16 (SD = .17; range = 0-.84)
Effect size of the moderation effect (495 out of 656 effects)	M = .17; M_w = .08; Mdn = .15 (SD = .13; range = 0-.87)
Correlation X with Z (150 out of 247 effects)	M = .17; M_w = .15; Mdn = .10 (SD = .16; range = 0-.67)
Explanatory variables (X or Z)	1,312
Manipulated X or Z (1,312 out of 1,312 variables)	390 (30%)
Measured X or Z (1,312 out of 1,312 variables)	922 (70%)
Measure reliability of X or Z (767 out of 922 variables)	M = .88; M_w = .86; Mdn = .87 (SD = .10; range = .45-.99)
Number of indicators of X or Z (914 out of 922 variables)	M = 6.71; Mdn = 4; Mo = 3 (SD = 10.61; range = 1-169)
Continuous x or z (919 out of 922 variables)	120 (13%)
Categorical x or z (919 out of 922 variables)	799 (87%)
Number of scale points of x or z (779 out of 799 variables)	M = 7.67; Mdn = 7; Mo = 7 (SD = 10.72; range = 2-101)
Outcome variables (Y)	656
Measure reliability of Y (266 out of 656 variables)	M = .90; M_w = .87; Mdn = .90 (SD = .08; range = .51-.98)
Number of indicators of Y (504 out of 656 variables)	M = 2.41; Mdn = 1; Mo = 1 (SD = 2.14; range = 1-13)
Continuous y (653 out of 656 variables)	208 (32%)
Categorical y (653 out of 656 variables)	445 (68%)
Number of scale points of y (430 out of 445 variables)	M = 11.47; Mdn = 7; Mo = 7 (SD = 22.5; range = 2-200)

Notes: Literature review of moderation analyses in the 2015-2019 volumes of *Journal of Marketing Research* (*JMR*), *Journal of Marketing* (*JM*), *Journal of Consumer Research* (*JCR*) and *Marketing Science* (*Mark. Sci.*). Percentages may not sum to 100% due to rounding or use of multiple methods within an article. The numbers in parentheses in the first column denote the number of articles, studies, effects, or variables that the corresponding statistics in the remaining columns are based on. Discrepancies manifest because some statistics could not always be unequivocally determined from study descriptions. Effect sizes are correlations (r) and are based on absolute values of zero-order correlations, and on transformed test-statistics (t, z, χ^2 or F) if zero-order correlations were not available (Rosenthal and DiMatteo 2001). Means of effect sizes, correlations and reliabilities are based on Fisher-Z-transformed values. Weighted means use $\sqrt{n-3}$ weights, the inverse standard error of Fisher-Z (where n is the sample size). M is the mean, M_w is the weighted mean, Mdn is the median, Mo is the mode, and SD is the standard deviation.

In our sample, 637 (97%) out of the 656 interaction effects tested moderation hypotheses (i.e., the effect of XZ). The remaining 19 (3%) tested U-shape (e.g., a quadratic effect X^2) hypotheses (Haans et al. 2016). Thus, U-shapes are relatively rare. Moderation effects with measured variables were common, 266 (41%) out of 656 effects. The remainder had moderation of a measured variable with a manipulation (59%).

We also document the size of the moderation effects, as well as properties of the measures. Reported zero-order correlations determined effect sizes for main effects. Meta-analysis estimated the mean reliabilities and correlations. We transformed reliabilities and correlations to Fisher-Z-values, took the mean, and back-transformed it to a meta-analytic mean correlation or reliability. Table WA2 reports simple and weighted means, which use the inverse of the standard error of the Z-values $\sqrt{n-3}$ as weights, where n is the sample size, giving more weight to observations from larger studies. The mean absolute effect size is r = .20, which is a small-to-medium effect (Cohen 1988) and slightly lower than the mean effect size of .24 found in a meta-analysis of meta-analyses in marketing (Eisend 2015). Because correlations between interaction terms and Y are uncommonly reported, we transformed exact t- z- χ^2- and F-statistics when correlations were unavailable (Rosenthal and DiMatteo 2001). The mean effect size is r = .17, again a small-to-medium effect (Cohen 1988). The magnitudes are consistent with the conventional wisdom that moderation effect sizes are smaller than main effects (Aguinis et al. 2005; Eisend 2015).

The measured explanatory variables (922 out of 1,312 X and Z variables) had a mean reliability of .88, which is good to excellent (Peterson 1994) and in line with Pieters (2017), but substantively higher than the mean of .77 found in an early meta-analysis (Peterson 1994). Three indicators were most common, with a mean 6.71 and a median number of four. Categorical indicators (87%) most commonly used seven-point scales (62% at the mode). The outcome variable Y had an excellent mean reliability of .90 (Peterson 1994). It had a

mean 2.41 indicators, similar to the mean of 2.28 found by Pieters (2017), but was most commonly measured with a single-indicator (54%). Outcomes were usually categorical (445 out of 656 variables, 68%). These measures most commonly used seven-point scales (53% at the mode).

Out of the 1,003 non-manipulated, multi-indicator Y, X and Z measures for which information on the measurement could be unequivocally determined, 822 (82%) reported a Cronbach's alpha (or a correlation that we transformed to an alpha estimate) without specific tests whether indicators are equally good, and 948 (95%) used equal weighting to create a mean composite. The remainder used composite reliability estimates that account for unequal weighting (Fornell and Larcker 1981; Raykov 1997).

Web Appendix B: Example code

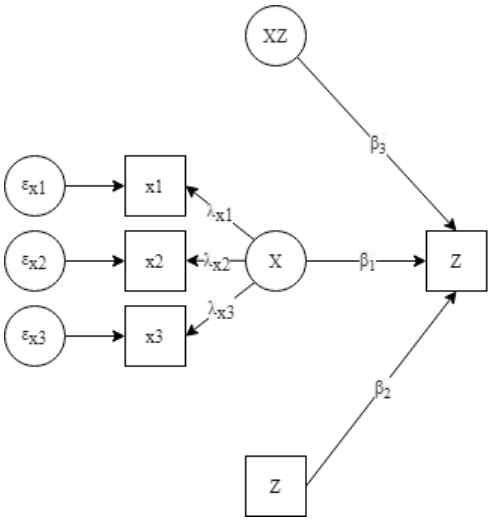
Table WA3 has an overview of example code to implement the methods, available on OSF:

https://osf.io/py7jx/?view_only=5d921a6658cf402a80bd1d4996665331.

Table WA3
Overview of Example Code

Model	Method	Implementation				
		SPSS	*Stata*	*R-lavaan*	*R-nlsem*	*Mplus*
Model 1: Latent X with 3 continuous indicators Latent Z with 3 continuous indicators Latent Y with 3 continuous indicators	1. Multi-group	×	✓	✓	×	✓
	2. Means	✓	✓	✓	×	✓
	3. Corrected means	×	✓	✓	×	✓
	4. Factor scores	×	✓	✓	×	×
	5. Product indicators	×	✓	✓	×	✓
	6. Latent product	×	×	×	✓	✓
Model 2: Latent X with 3 continuous indicators Manifest discrete (binary) Z Latent Y with 3 continuous indicators	1. Multi-group	×	✓	✓	×	✓
	2. Means	✓	✓	✓	×	✓
	3. Corrected means	×	✓	✓	×	✓
	4. Factor scores	✓	✓	✓	×	×
	5. Product indicators	×	✓	✓	×	✓
	6. Latent product	×	×	×	×	✓

Table WA3 (CONTINUED)

Model	Method	Implementation				
		SPSS	*Stata*	*R-lavaan*	*R-nlsem*	*Mplus*
Model 3: Latent X with 3 continuous indicators Manifest discrete (binary) Z Manifest continuous Y  	1. Multi-group	×	✓	✓	×	✓
	2. Means	✓	✓	✓	×	✓
	3. Corrected means	×	✓	✓	×	✓
	4. Factor scores	✓	✓	✓	×	✓
	5. Product indicators	×	✓	✓	×	✓
	6. Latent product	×	×	×	×	✓

Note: Method availability and non-availability in statistical software are denoted by ✓ and × respectively. Available code can be accessed on OSF: https://osf.io/py7jx/?view_only=5d921a6658cf402a80bd1d4996665331.

Web Appendix C: Factor scores method

Factor score estimators

Latent variables are unobservable, but they can be estimated with factor scores. There are multiple ways to estimate factor scores. Yet, the choice of factor scores matters because of their different properties. Three types of factor scores are dominant (DiStefano et al. 2009; Lastovicka and Thamodaran 1991). A first factor score estimator is the regression method (Lastovicka and Thamodaran 1991, Equation 5). In regression terminology, the dependent variable is the factor score to be estimated, the independent variable is the matrix of observed indicator data and the regression parameters are the estimated correlation between the indicators and the latent factors. Thus, the regression factor scores minimizes the sum of squares of factor scores with the true scores; or in other words, it maximizes the estimated correlation between the indicators and the factors. It essentially optimally weights the indicator data by the loading weights and does not use the measurement error estimates (see Bartlett scores below). The variance of regression factor scores is therefore the estimated proportion of variance extracted by the factor from the items, which is equal to the estimated reliability or ρ if the variance of the factor is one (Yuan et al. 2020, p. 338). Formally:

$$\hat{F}_{Regression} = D\Sigma_{(o)}^{-1}\Lambda\Phi, \quad \text{(W1)}$$

where D is a matrix of indicator-level data that is multiplied by the inverse of the observed covariance matrix $\Sigma_{(o)}^{-1}$, the matrix of estimated loadings Λ and the estimated variance covariance matrix of the latent variables Φ.

A second factor score estimator is the Bartlett method (Lastovicka and Thamodaran 1991, Equation 3) that is similarly based on a regression solution, but weighted by the measurement error variances. Holding the loadings equal, it weights indicators with a small amount of measurement error more than it weights indicators with a large amount of measurement error. Formally:

(W2) $$\hat{F}_{Bartlett} = D\Theta^{-2}\Lambda(\Lambda^T\Theta^{-2}\Lambda)^{-1},$$

This estimator minimizes the sum of squares for the unique factors. It is the optimal linear combination of the indicators with maximal reliability, for instance in the case of three indicators for X, for observation i:

(W3) $$\hat{F}_{Bartlett,i} = \frac{\lambda_{x1}}{\sigma_{x1,x1}}x_{1,i} + \frac{\lambda_{x2}}{\sigma_{x2,x2}}x_{2,i} + \frac{\lambda_{x3}}{\sigma_{x3,x3}}x_{3,i}.$$

Hence, the factor score variance is ϕ/ρ, or the inverse of the factor reliability if the variance of the factor is one such that Bartlett scores have the same scale as mean scores (Yuan and Deng 2021, Equation 7).

A third factor scores estimator is the Anderson-Rubin method (Lastovicka and Thamodaran 1991, Equation 6). It builds on the Bartlett estimator to obtain factor scores that are uncorrelated and standardized (Lastovicka and Thamodaran 1991). It is:

(W4) $$\hat{F}_{Anderson-Rubin} = D\Theta^{-2}\Lambda(\Lambda^T\Theta^{-2}\Sigma_{(o)}\Theta^{-2}\Lambda)^{-1/2}.$$

Factor score regression with Bartlett (Y) and regression (X & Z) scores

As explained in the main text, the challenge is to use the indicators to estimate the main and moderation effects of X and Z. Bias in the variance of Y due to measurement error does not directly bias estimates but underestimates the coefficient of determination R^2 (Wooldridge 2015). Because the factor score estimators presented above have different properties, their estimates of the moderation effects also differ. Anderson-Rubin scores cannot be used for factor score regression because the factor scores are orthogonal and can therefore not recover non-zero covariances between Y and the predictors (Lastovicka and Thamodaran 1991). However, we implement the results from Skrondal and Laake (2001) and Devlieger et al. (2016) in the latent moderation setting. These studies have shown that regression parameters in non-moderation models can be recovered if Bartlett factor scores are used for the outcome

variable (here $\hat{Y}$) and regression factor scores from a 2-CFA are used for the predictors (here $\hat{X}$ and $\hat{Z}$). The main text elaborates on this.

Using Bartlett scores corrects for unreliability in Y. The use of regression scores corrects for unreliability in the predictors. Therefore, the properties of both factor score estimators are combined to account for unreliability in all variables. Like in the main text, assuming that Y, X and Z are normally distributed, that X and Z are uncorrelated and that Bartlett factor scores are used for Y and regression scores for X and Z (and analogous for β_1 and β_2):

(W5) $$\hat{\beta}_{3,Bartlett-regression} = \frac{\rho_{XZ} * \phi_{Y,XZ}}{\rho_{XZ} * \phi_{XZ,XZ}} = \beta_3.$$

here, both the variance of XZ and its covariance with Y are attenuated by ρ_{XZ}, which cancels out to estimate β_3 (Devlieger et al. 2016). Importantly, the variances and covariances between the factor scores cannot be interpreted as estimates of the true variances and covariances (i.e., free of the impact of measurement error), but the regression estimate β_3 can.

However, if Bartlett scores are used for both Y and XZ (Devlieger et al. 2016):

(W6) $$\hat{\beta}_{3,Bartlett-Bartlett} = \frac{\phi_{Y,XZ}}{\phi_{XZ,XZ} * \frac{1}{\rho_{XZ}}} = \beta_3 * \rho_{XZ},$$

where the covariance between Y and XZ is correctly estimated but the variance of XZ is inflated by the inverse of the reliability. Thus, Bartlett scores only account for unreliability when used as a dependent variable. Similarly, if regression scores are used for both Y and XZ:

(W7) $$\hat{\beta}_{3,regression-regression} = \frac{\phi_{Y,XZ} * \rho_{XZ} * \rho_Y}{\phi_{XZ,XZ} * \rho_{XZ}} = \beta_3 * \rho_Y,$$

where the covariance between Y and XZ is attenuated by the reliabilities of both Y and XZ, and the variance of XZ is attenuated by ρ. Thus, regression scores only account for

unreliability when used as an predictor. In sum, the true main effects and the moderation effect are only recovered when the indicators do not contain measurement error (i.e., all ρ = 1) or when Bartlett scores for the outcome and regression scores for the predictors are used.

Factor score (moderated) mediation

One limitation of factor score regression with Bartlett scores for Y and regression factor scores for the predictors is that it requires determining a priori which of the variables are outcomes and which are predictors (Devlieger et al. 2016, pp. 747 & 763). One situation in which this might occur is a theory of (moderated) mediation. In that case, the mediator M is both an outcome and a predictor (in the equation for Y).

The mediation model, omitting intercepts and moderation for brevity, is:

(W8) $$M = a * X + \zeta_M,$$

(W9) $$Y = b * M + cp * X + \zeta_Y,$$

with $\zeta_M \sim N\left(0, \sigma^2_{\zeta_M}\right)$ and $\zeta_Y \sim N\left(0, \sigma^2_{\zeta_Y}\right)$. Commonly, the analyst is interested in a decomposition of the total effect of X on Y $c = a * b + cp$ in an indirect (or mediation) effect $a * b$ and a conditional (on M) direct effect cp (Pieters 2017). Thus, if a, b and cp can be estimated accurately, the focal total, indirect and direct effects of X on Y are also accurately estimated. A possible solution is to estimate both Bartlett and regression factor scores for M. For brevity, we focus on a non-moderated mediation model, but the results extend to moderated mediation.

For the M-equation, using Bartlett factor scores for M and regression scores for X accurately estimates a. For the Y-equation, Bartlett factor scores for Y and regression scores from a 2CFA of M and X accurately estimate b and cp. The factor scores mediation model, using subscripts to indicate the type of factor scores, becomes:

(W10) $$M_{Bartlett} = a * X_{Regression,factorwise} + \zeta_M,$$

(W11) $$Y_{Bartlett} = b * M_{Regression,blockwise} + cp * X_{Regression,blockwise} + \zeta_Y.$$

Step 1 then estimates three factor analyses. Factor analysis one is a factor analysis of M (1-CFA) on its indicators of which Bartlett scores are estimated. These factor scores are entered as the outcome in the M-equation. Factor analysis two is a factor analysis of X (1-CFA) on its indicators from which factorwise regression factor scores are estimated; these serve as the predictor in the M-equation. Factor analysis three is a 2-CFA (blockwise) of M and X on their respective indicators because they enter both in the Y-equation. The regression factor scores of M and X enter as predictors in Equation (W11). Step 2 simultaneously or separately estimates the two factor score regressions.

The code available on OSF gives an example of this estimation technique for a moderated mediation model in R (R Core Team 2020) with the factor scores method. For comparison, it also implements the moderated mediation model in Mplus (latent product method) that estimates the measurement models with the structural mediation model simultaneously.

Web Appendix D: Non-normality

The latent product method estimates the moderation effect by using the latent product of X and Z (Klein and Moosbrugger 2000). It is a distribution analytic approach, i.e., based on an analysis of the indicator distribution (Kelava et al. 2011) instead of the covariance matrix, like the other methods. The latent product method is motivated by the finding that Y and its indicators are non-normally distributed if there is a true moderation effect, even if X and Z are normally distributed (Kenny and Judd 1984; Klein and Moosbrugger 2000; Moosbrugger et al. 1997). None of the methods except for the latent product method account for this feature in the data. The latent product method was developed to take the non-normality in Y into account, while maintaining the assumption of normally distributed indicators of X and Z (Klein and Moosbrugger 2000).

The non-normality in moderation models follows from the properties of products of distributions. That is, even if X and Z are normally distributed, a product of normally distributed variables is usually not normally distributed (Aroian 1947; Oliveira et al. 2016). The univariate skewness and excess kurtosis of the distribution are conventional metrics for the degree of (non-)normality of a distribution. Both are zero for normally distributed variables. Skewness implies asymmetry of the distribution. For example, negative skewness in a satisfaction distribution reflects that customers are generally satisfied and not dissatisfied (vice versa for positive skewness) with the products they purchase and consume (Peterson and Wilson 1992). A positive (negative) excess kurtosis reflects a higher (lower) likelihood that there are extreme observations in the tails of the distribution than there would be in a normal distribution.

To visualize this non-normal distribution, Panel A in Figure WA1 (code is available on OSF) visualizes (solid line) a simulated (n = 100,000) density of $X * Z$ with X and Z being standard normally distributed with a .20 correlation. The dashed line represents the normal

distribution with the same mean and standard deviation as $X * Z$. The figure shows that the product of X and Z is skewed (estimated skewness about 1.13) and has excess kurtosis (estimated kurtosis about 6.57), even though both X and Z are standard normally distributed (skewness and excess kurtosis both zero).

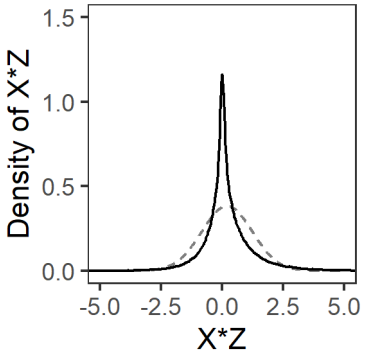
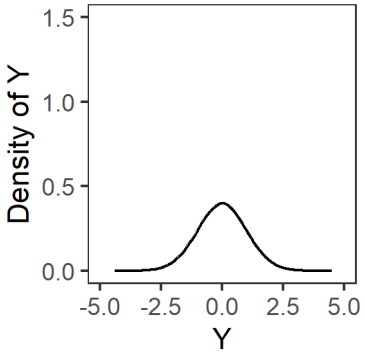
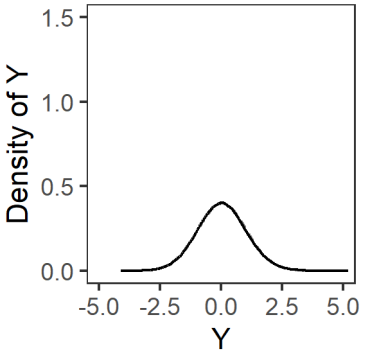
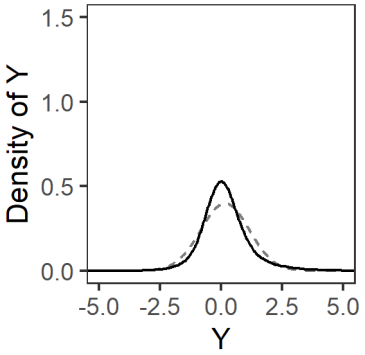
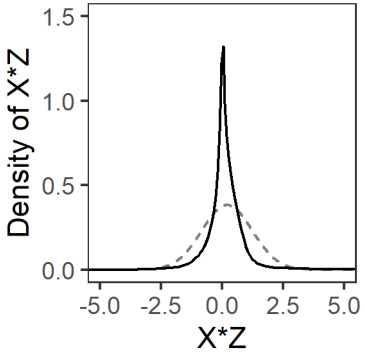
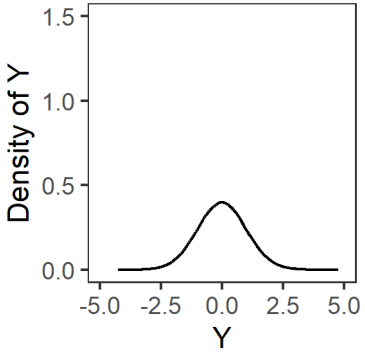
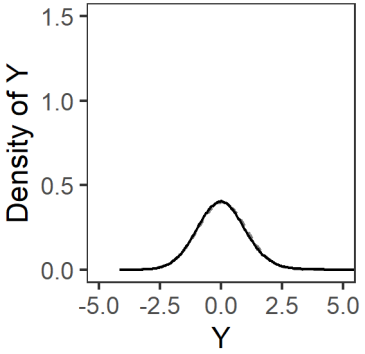
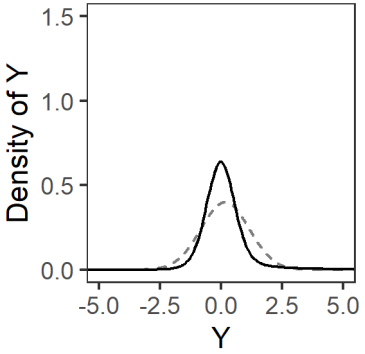
The degree of non-normality of a product of normal distributions is a complex function of the means and standard deviations of the distributions, and the correlations between them (Oliveira et al. 2016, Equations 3-6). To illustrate the non-normality of product terms in the current setting, we focus on the situation where X and Z are standard normally distributed. The mean of $X * Z$ is $\phi_{X,Z}$ (the correlation of X with Z), the variance is $1 + \phi_{X,Z}^2$ and the skewness and excess kurtosis are (Oliveira et al. 2016, Equations 3-6):

(W12) $$Skew_{XZ} = \frac{2*\phi_{X,Z}*(3+\phi_{X,Z}^2)}{(1+\phi_{X,Z}^2)^{2/3}},$$

(W13) $$Kurt_{XZ} = \frac{6*(1+6*\phi_{X,Z}^2+\phi_{X,Z}^4)}{(1+\phi_{X,Z}^2)^2}.$$

Figure WA2 plots the skewness and kurtosis of XZ as a function of $\phi_{X,Z}^2$ (code is available on OSF). Both are positive and increase when the absolute value of the correlation increases. When the correlation is one, the skewness is about five, while the excess kurtosis is nine. Even when X and Z are uncorrelated, the excess kurtosis is three, the minimum, while the skewness is zero. Thus, even when standard normally distributed X and Z are uncorrelated, their product is non-normally distributed. If they are correlated, the extent of non-normality is stronger for a higher correlation of X with Z.

Figure WA1
Y is Non-Normally Distributed if There is a True Moderation Effect

	Panel A: density of X*Z for normal X & Z	Panel B: density of Y for normal X & Z ($\beta_3 = 0$)	Panel C: density of Y for normal X & Z ($\beta_3 = .20$)	Panel D: density of Y for normal X & Z ($\beta_3 = .80$)
Normal X & Z (skewness = 0 & excess kurtosis = 0)				
	Panel E: density of X*Z for non-normal X & Z	Panel F: density of Y for non-normal X & Z ($\beta_3 = 0$)	Panel G: density of Y for non-normal X & Z ($\beta_3 = .20$)	Panel H: density of Y for non-normal X & Z ($\beta_3 = .80$)
Non-normal X & Z (skewness = 2 & excess kurtosis = 7)				

Notes: Solid lines are densities of X*Z or Y, where X and Z are simulated (n = 100,000) standard normally distributed with a .20 correlation and main effects β_1 and β_2 are .20. Dashed lines are densities of the normal distribution with the same mean and standard deviation of X*Z or Y.

Because Y (Equation 1 in the main text) is a weighted (by β) function of two normally distributed variables (X and Z) and one typically non-normally distributed variable ($X * Z$), Y is also non-normally distributed if there is a true moderation effect. The extent of non-normality in Y depends on the non-normality of the product but also the strength of the moderation effect. To demonstrate this, Panels B to D in Figure WA1 build on Panel A. If the true moderation effect is zero (Panel B), or .20 (Panel C), about average (see Table 1 in the main text), the distribution of Y remains approximately normally distributed. However, if the moderation effect is very large, here .80 (Panel D), the distribution of Y also becomes more non-normal (estimated skewness and kurtosis are respectively about .91 and 3.93 here). In sum, Y and its indicators are also non-normally distributed if there is a non-zero moderation effect and even if X, Z and ζ are normally distributed. Thus, if there is a true moderation effect:

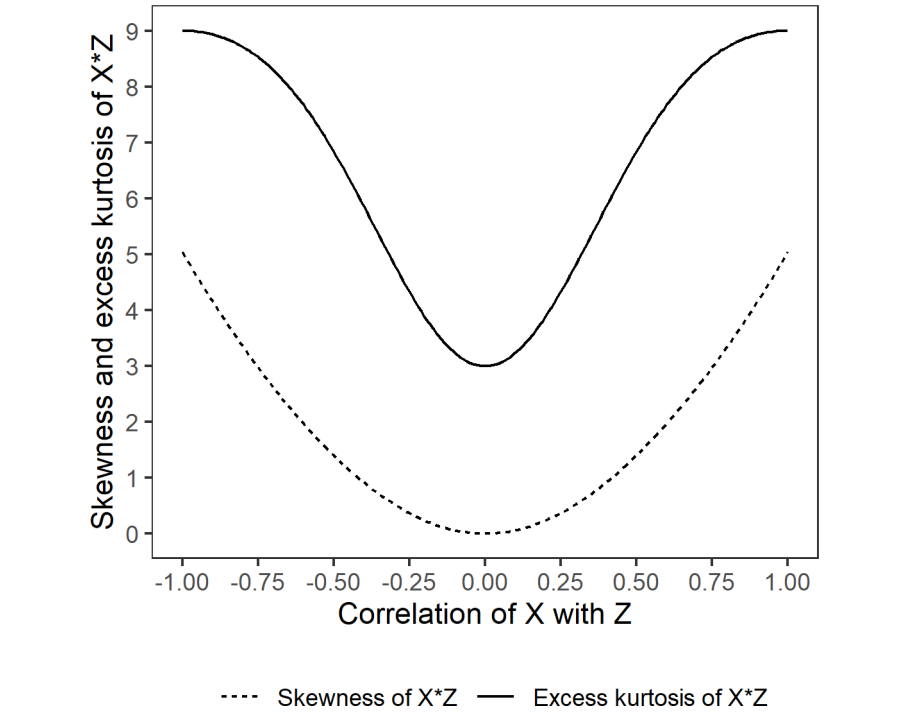
(W14) $$Y \nsim MVN(\mu, \Sigma).$$

Non-normality in X and Z is also further exacerbated by the product term. To illustrate this, we generate data under the same conditions as before, but now with X and Z having skewness of 2 and excess kurtosis of 7, which represent levels of non-normality that lead to issues with maximum likelihood estimation such as parameter and standard error bias (Finney and DiStefano 2006, p. 442). The distribution of $X * Z$ (Panel E in Figure WA1) is even more peaked (estimated skewness is about 7 and excess kurtosis almost 100) than that in Panel A, which also results in a more non-normal Y the stronger the moderation effect becomes (Panels F-H). In sum, non-normality in Y is exacerbated by non-normality in X and Z (and if non-normal, for stronger main effects β_1 and β_2), for stronger moderation effects (β_3) and for higher correlations between X and Z.

Nevertheless, the multi-group, corrected means, factor scores and product indicators methods use a measurement model for Y and therefore assume that Y is normally distributed

(Bollen 1989; Finney and DiStefano 2006, p. 441). Previous research concluded that skewness and excess kurtosis in variables lead to overestimated zero-order correlations (Bishara and Hittner 2015). We would therefore expect inflated main and moderation effects. Larger levels of kurtosis might also bias standard errors downward (Finney and DiStefano 2006, p. 444). It is however unclear to what extent this leads to bias in settings that are common in the literature review. For instance, effect sizes are commonly small-to-medium, about .20 or smaller (see Table 1 in the main text). Study 1 in the main text investigates this.

Figure WA2
Skewness and Kurtosis of a Product of Standard Normally Distributed Variables Increase if their Correlation Increases



Web Appendix E: Latent product method

Mixture approximation

The latent product method accounts for the non-normality in Y and its indicators by directly fitting the multivariate distribution of the indicators instead of their covariance matrix (Klein and Moosbrugger 2000). The challenge of approximating the indicator distribution is to take the non-normality due to the interaction into account. Klein and Moosbrugger (2000) proposed that the non-normal distribution of Y can be approximated by a finite mixture of normally distributed variables. The finite mixture is a tool to approximate the non-normal distribution with multiple tractable normal distributions. The number of distributions or mixture components must be fixed prior to estimation and represents a tradeoff between accuracy and computational intensiveness.

Figure WA3 illustrates the use of a mixture of normal distributions to approximate the non-normal distribution of Y. Code is available on OSF. Panel A visualizes a simulated (n = 100,000) density of Y with X and Z standard normally distributed with a .50 correlation, main effects of .20 and a moderation effect of .80. Panel B approximates this non-normal density (solid line) with a single normal distribution (dashed line) that has the same mean and standard deviation. Clearly, and intuitively, the normal distribution inaccurately approximates the non-normal distribution of Y: the non-normal distribution is more peaked and skewed to the left than the normal distribution is.

Panel C visualizes two normally distributed components to approximate Y and Panel D has the mixture density, the sum of both densities. Using a mixture density with two components already fits the non-normal Y better, although some deviations persist, like in the peak and right tail of the distribution. Panels E and F visualize a four-component mixture distribution. Here, the joint mixture density (dashed line in Panel F) almost perfectly overlaps the distribution of Y. Generally, the more components, the better the distribution of Y can be

approximated, which fosters estimation accuracy but is more computationally intensive. Klein and Moosbrugger (2000, p. 465) recommend 16 mixture components for adequate approximation of the non-normal indicator distribution in moderation analyses with a single interaction, as is the focal case here. A follow-up study further explores this.

In sum, the key idea of the latent product method is to directly approximate the non-normal Y distribution due to the interaction. It uses a mixture of normal distributions to do so. We now turn to the mathematical background.

Model specification

The derivation of the result that the non-normal indicator distribution can be represented by a mixture distribution is based on the finding that the distribution is multivariate normal if it is conditioned on the components of the interaction (Moosbrugger et al. 1997). To illustrate this, rearrange Equation (1) in the main text:

(W15) $$Y = (\beta_1 + \beta_3 * Z) * X + \beta_2 * Z + \zeta.$$

When holding X at a fixed value (and analogous for Z), the effect of Z on Y is a linear function of Z, such that Y is a (weighted, by β) sum of normally distributed variables only and therefore also normally distributed (Moosbrugger et al. 1997).

Klein and Moosbrugger (2000) show that if a vector c is used as a conditioning variable, which is a rescaled (by ϕ) vector of standard normal distributions to represent X and Z, the conditional indicator distribution is normally distributed:

(W16) $$(y, x, z|c) \sim MVN(\mu_{(i)}, \Sigma_{(i)}),$$

where the implied (i) mean vector μ and covariance matrix Σ are complex functions of c and the model to be estimated in Equations (1)-(2) in the main text. They are provided by Equations (18)-(22) in Klein and Moosbrugger (2000) and by Equations (8)-(12) in Schermelleh-Engel et al. (1998) for the special case of an observed Y and two indicators for X and Z.

Using the product rule of conditional distributions, the indicator distribution can then be represented as the product of the distributions of c and the conditional distribution (Klein & Moosbrugger 2000, Equation 15):

(W17) $$f(y = y, x = x, z = z) = \int \varphi(0,1)(c)\varphi(\mu_{(i)}, \Sigma_{(i)})(y, x, z)\mathrm{d}c,$$

where φ is the standardized normal density. The integral in Equation (W17) represents a mixture of normal densities (Klein and Moosbrugger 2000).

To summarize, the latent product method is a distribution analytic approach, a direct analysis of the indicator distribution of Y, X and Z. Instead of the observed covariance matrix, it requires the raw data of the indicators. Although the distribution of Y is non-normally distributed if there is a true moderation effect, it is normally distributed when conditioning on X and Z. The latent product method utilizes this property to derive the distribution of the indicators.

Model estimation

Although the integral in Equation (W17) cannot be solved analytically (Klein and Moosbrugger 2000, p. 464), it can be approximated by a finite mixture of K normal densities. Here, the finite mixture is a weighted sum of normal distributions to approximate the non-normal indicator distribution. The weights are derived analytically and do not have to be estimated (Klein and Moosbrugger 2000). However, the number of mixture components K has to be fixed which represents a tradeoff between accuracy and computational intensiveness. Formally, the finite mixture distribution is (Klein & Moosbrugger 2000, Equation 29):

(W18) $$f(y = y, x = x, z = z) = \sum_{j=1}^{K} w_j \, \varphi(\mu_j, \Sigma_j)(y, x, z),$$

which is a weighted sum of K mixture components. Then ws are the analytically derived mixture weights (Klein and Moosbrugger 2000).

The log-likelihood is:

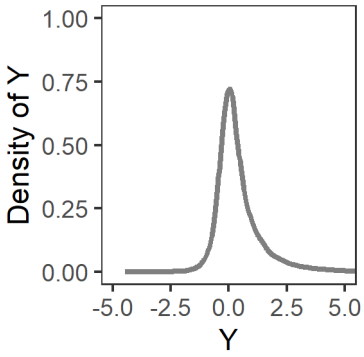
(W19) $$\ln L(\theta) = \sum_{i=1}^{N} \ln\left(\sum_{j=1}^{K} w_j \, \varphi\left(\mu_j, \Sigma_j\right)\right).$$

In mixture models, each observation i has a probability P_{ij} to belong to each (j) of the K mixture components (segments). These probabilities and the set of parameters to be estimated (θ) are not jointly identified but they can be estimated if the other one is known. Therefore, the latent product method uses an expectation maximization (EM) algorithm for estimation (Dempster et al. 1977; Klein and Moosbrugger 2000), which is common when estimating parameters with mixture densities. EM proceeds in two iterative steps (Dempster et al. 1977).

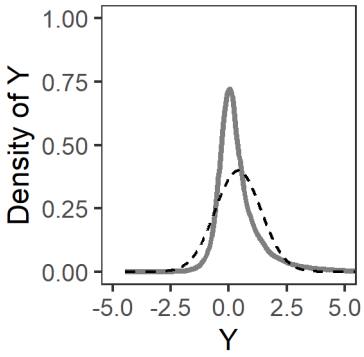
The *expectation* step of iteration r of the EM algorithm (for the first iteration, the parameter vector θ is initialized with starting values) calculates the probabilities of the mixture component *j* for observation i (Klein & Moosbrugger 2000, Equation 30). Once estimates of the mixture probabilities are available, the parameter vector $\theta^{(r)}$ can be updated with the likelihood in the *maximization* step (Klein & Moosbrugger 2000, Equation 31). The algorithm iterates back and forth between expectation (updating mixture probabilities based on the parameter estimates from the maximization step from the previous iteration) and maximization steps (updating the parameter estimates based on the posterior distribution of mixture probabilities) until convergence is attained (i.e., when the likelihood is maximized). Details are in Klein and Moosbrugger (2000) and Umbach et al. (2017). Henceforth, the algorithm converges to maximum likelihood estimates of the main and moderation effects (Dempster et al. 1977; Klein and Moosbrugger 2000).

Figure WA3
A Mixture of Normal Densities Approximates the Non-Normal Distribution of Y

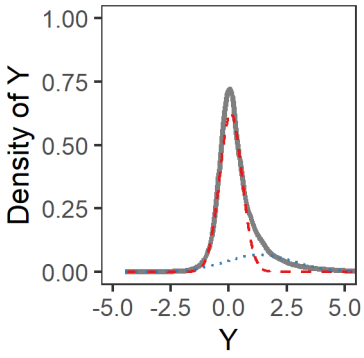
Panel A: density of Y



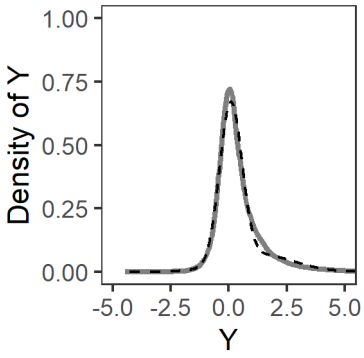
Panel B: density of Y and approximation by normal distribution



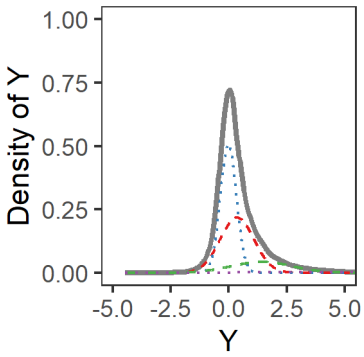
Panel C: density of Y and two normally distributed mixture components



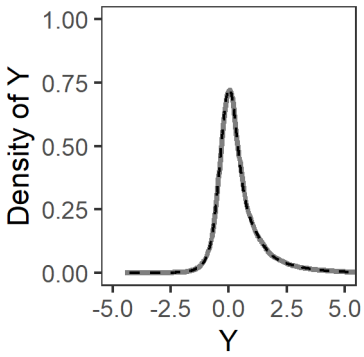
Panel D: density of Y and approximation by two normally distributed mixture components



Panel E: density of Y and four normally distributed mixture components



Panel F: density of Y and approximation by four normally distributed mixture components



Notes: Solid lines are densities of Y, where X and Z are simulated (n = 100,000) standard normally distributed with a .50 correlation, main effects β_1 and β_2 are .20. and moderation effect β_3 is .80. Panels C and E visualize normally distributed mixture components (non-solid lines) and panels B, D and F have the approximations of the density of Y of mixtures with one, two and four normally distributed mixture components.

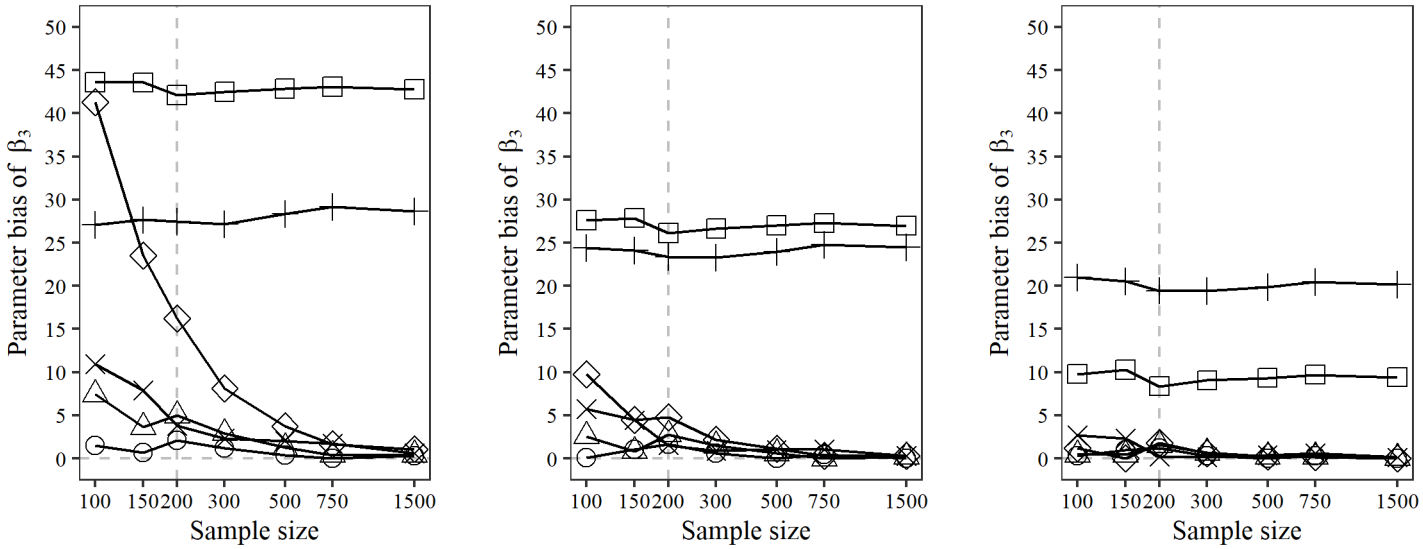
Web Appendix F: Study 1

Figure WA4 has performance criteria of the moderation effect for ten-fold cross validation.

The material on OSF plots the performance criteria for the main effects.

Figure WA4
Study 1: Ten-Fold Cross-Validation Performance Criteria for the Moderation Effect (β_3)

Panel A: Parameter bias of β_3



Panel B: Standard error bias of β_3

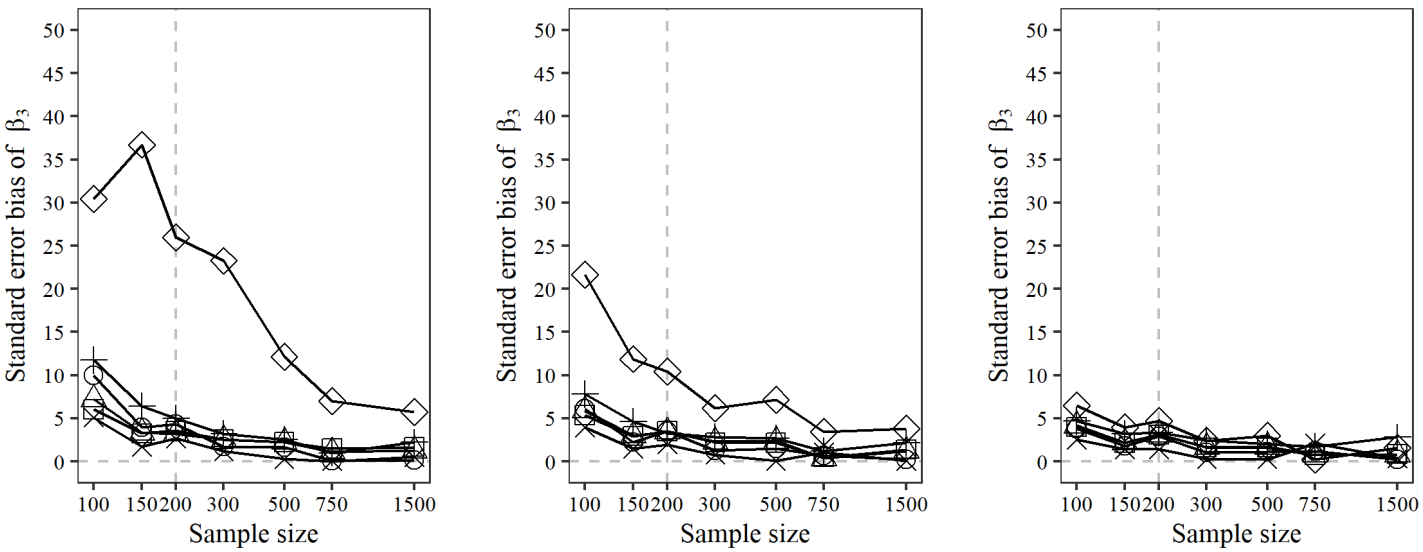
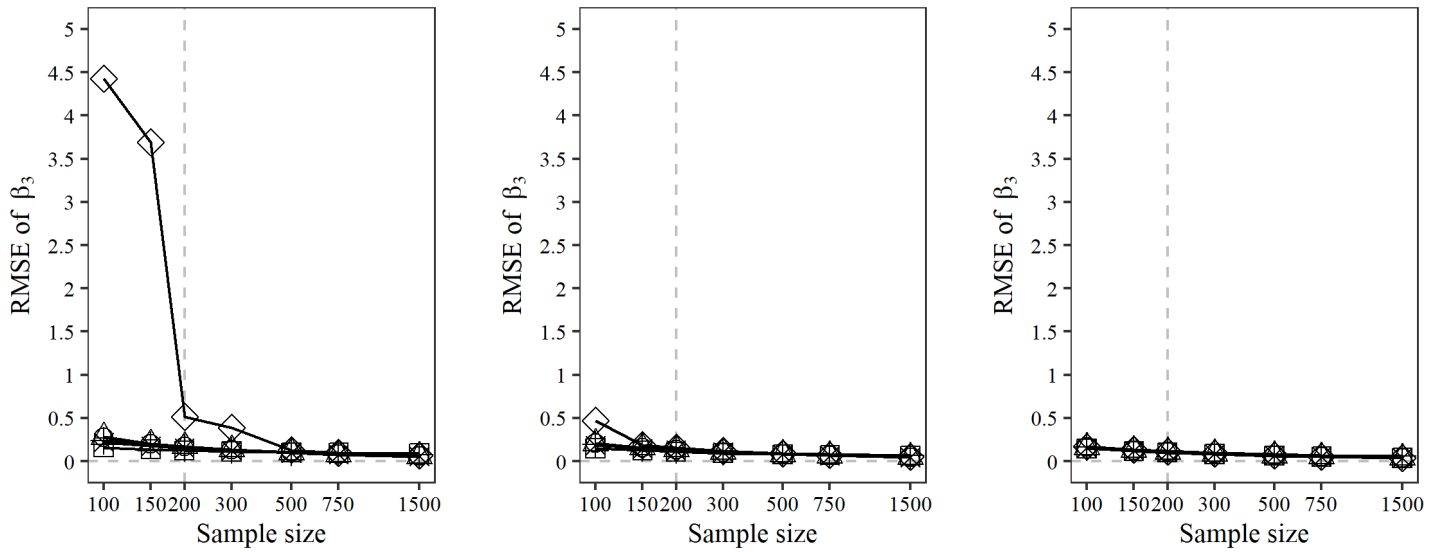


Figure WA4 (CONTINUED)

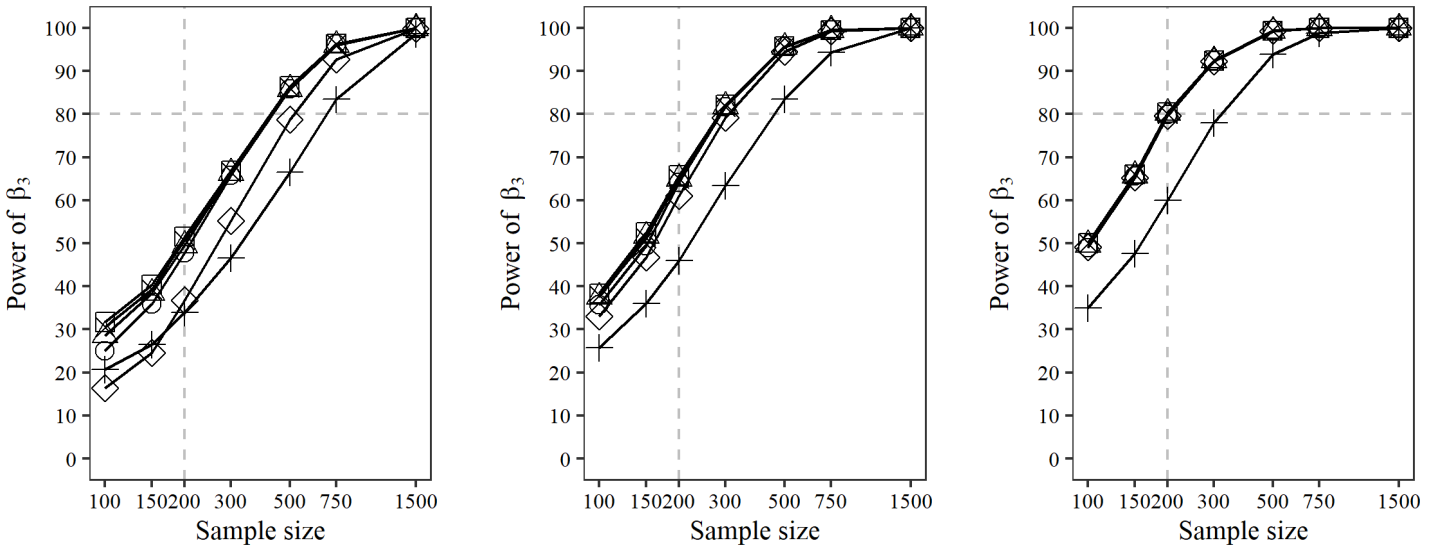
Panel C: RMSE of β_3

Reliability of Y, X and Z is .75 | Reliability of Y, X and Z is .85 | Reliability of Y, X and Z is .95



Panel D: Power of β_3

Reliability of Y, X and Z is .75 | Reliability of Y, X and Z is .85 | Reliability of Y, X and Z is .95



Legend:

1. Multi-group | 2. Means | 3. Corrected means
4. Factor scores | 5. Product indicators | 6. Latent product

Notes: Plots visualize method parameter bias, standard error bias, root mean squared error (RMSE) and power (as defined in Table 2 in the main text) of the moderation effect (β_3) across reliabilities of Y, X and Z. The sample sizes on the horizontal axes are on a log scale and indicate the sample sizes (90% for each fold) of the full estimation sample that is used to perform the ten-fold cross-validation. Horizontal dashed lines indicate parameter bias, standard error bias and RMSE of zero and power of 80%. Vertical dashed lines indicate a sample size of 200, about the median in the literature review (Table 1 in the main text).

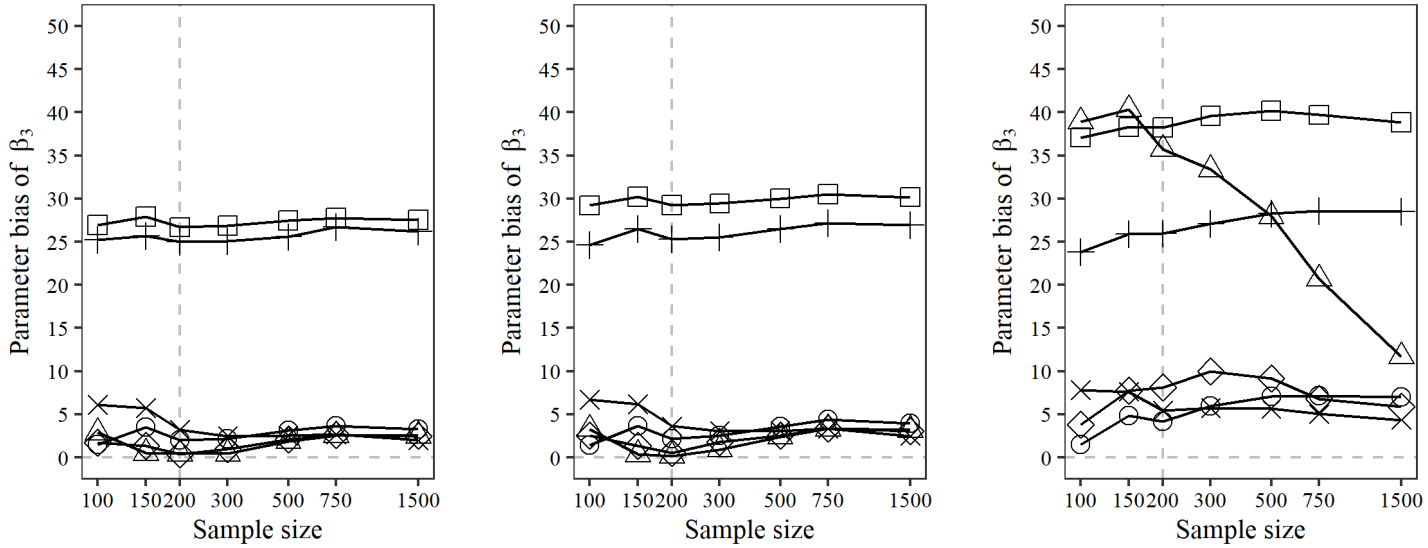
Web Appendix G: Study 2a

Figure WA5 has detailed results. OSF has the performance criteria for the main effects.

Figure WA5
Study 2a: Performance Criteria for the Moderation Effect (β_3)

Panel A: Parameter bias of β_3

Seven-point scales | Five-point scales | Three-point scales



Panel B: Standard error bias of β_3

Seven-point scales | Five-point scales | Three-point scales

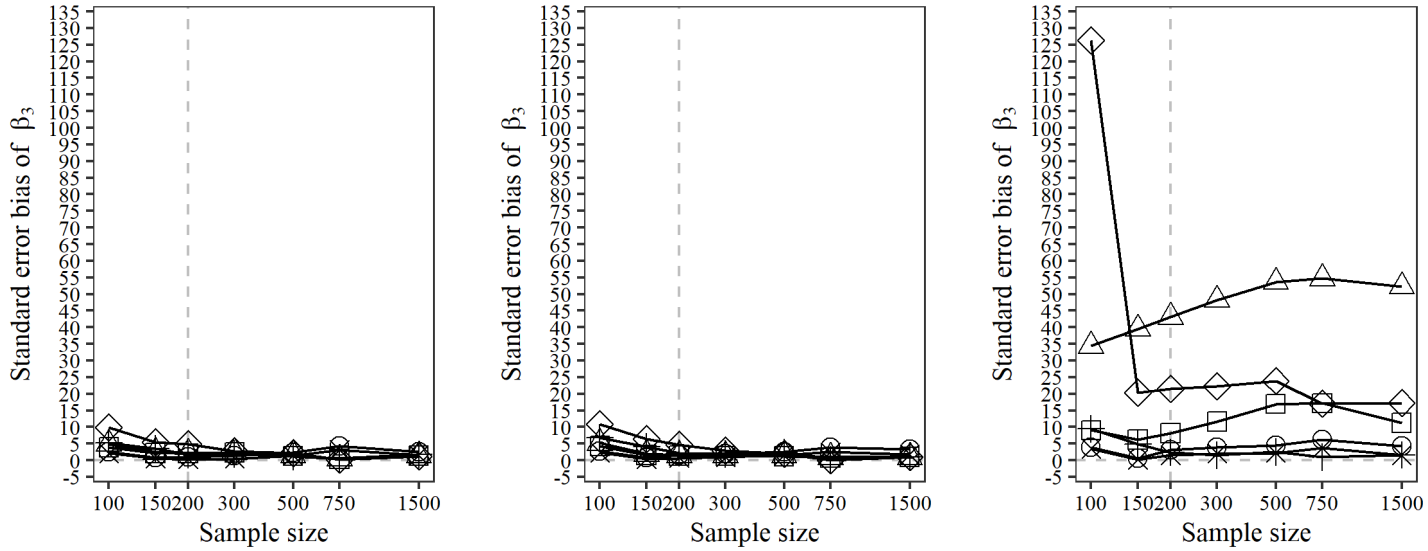
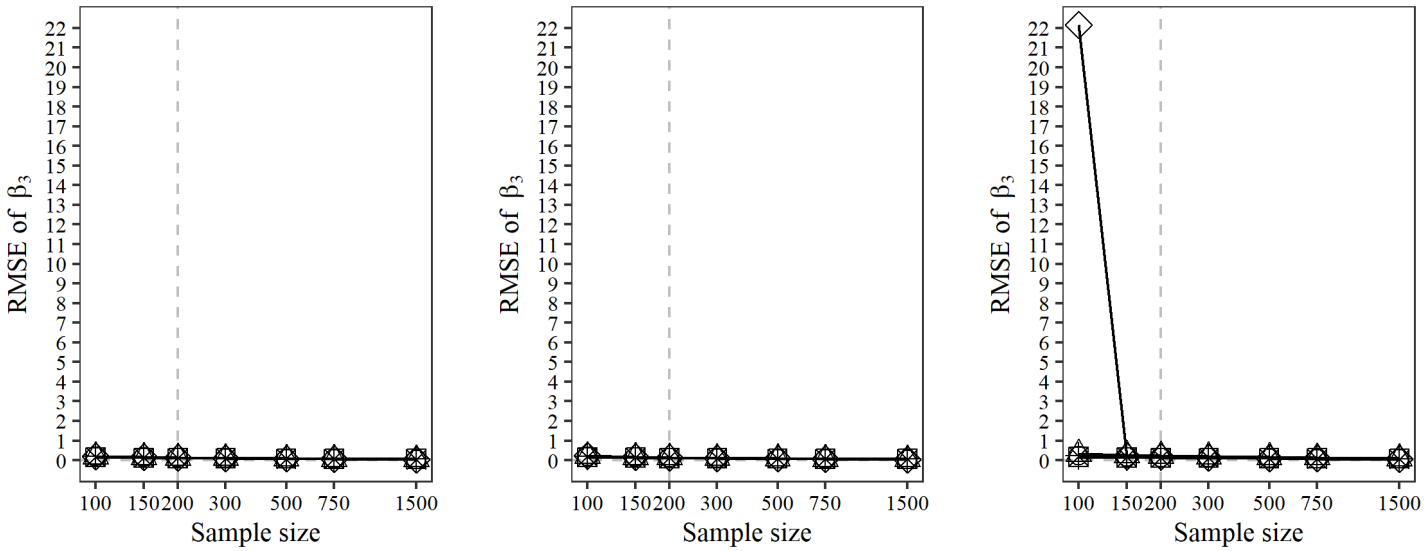


Figure WA5 (CONTINUED)

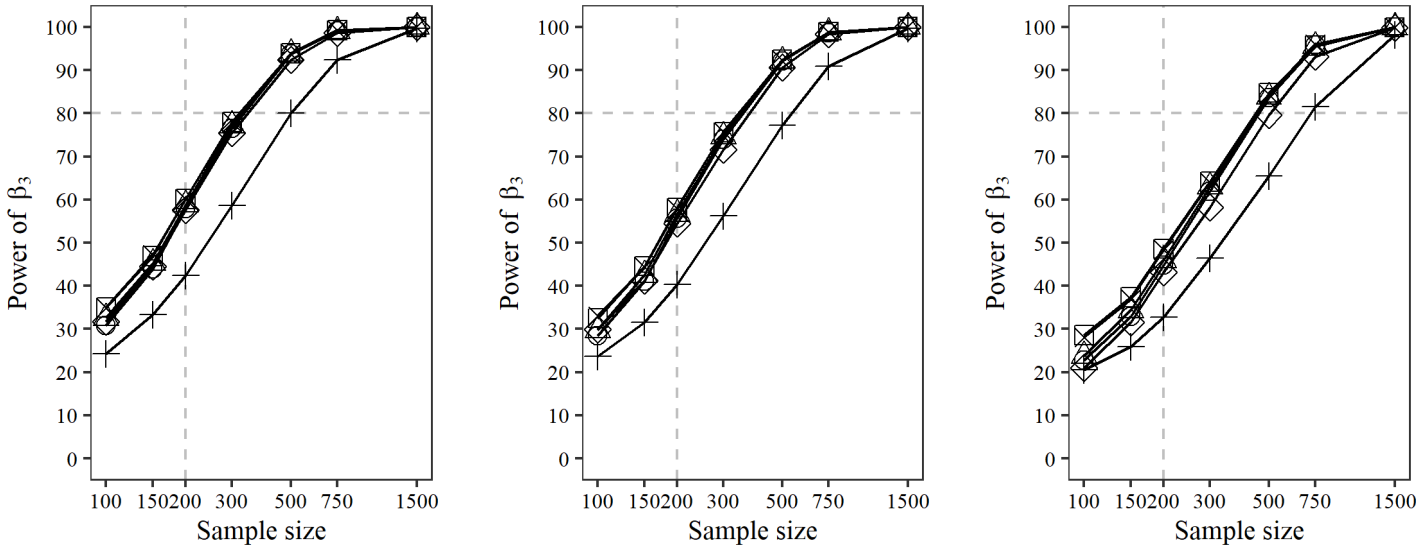
Panel C: RMSE of β_3

Seven-point scales | Five-point scales | Three-point scales



Panel D: Power of β_3

Seven-point scales | Five-point scales | Three-point scales



Legend:

1. Multi-group
2. Means
3. Corrected means
4. Factor scores
5. Product indicators
6. Latent product

Notes: Plots visualize method parameter bias, standard error bias, root mean squared error (RMSE) and power (as defined in Table 2 in the main text) of the moderation effect (β_3) across sample sizes (log scale) and the number of scale points. Horizontal dashed lines indicate parameter bias, standard error bias and RMSE of zero and power of 80%. Vertical dashed lines indicate a sample size of 200, about the median in the literature review (Table 1 in the main text).

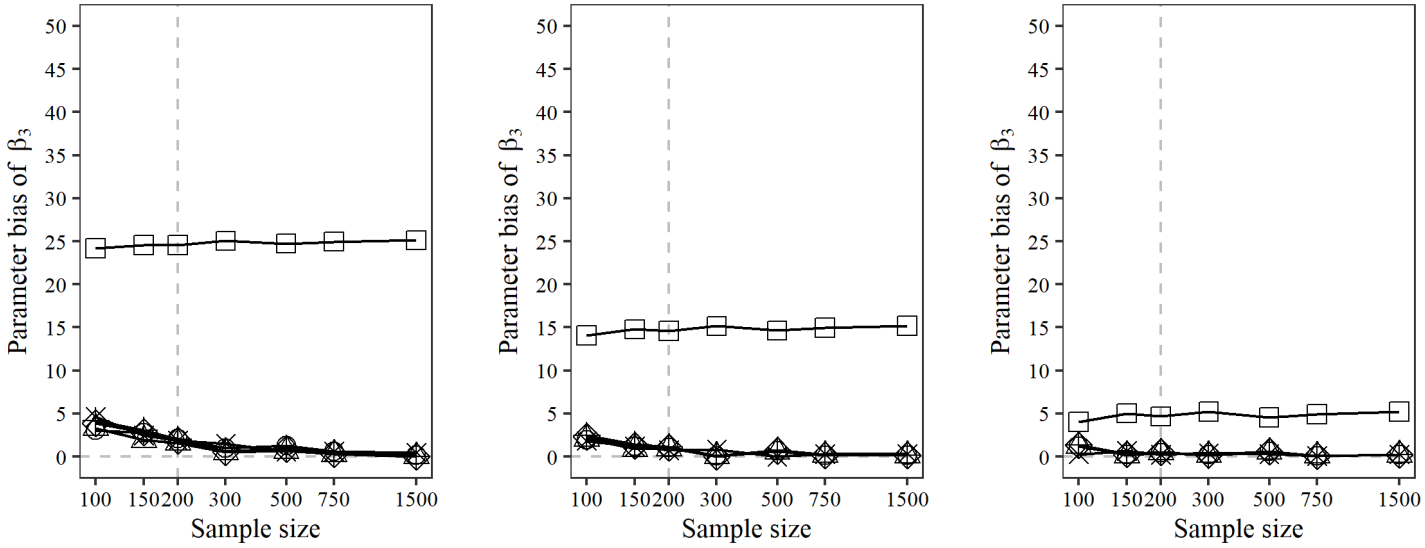
Web Appendix H: Study 2b

Figures WA6-8 have detailed results.

Figure WA6
Study 2b: Performance Criteria for the Moderation Effect (β_3)

Panel A: Parameter bias of β_3

Reliability of Y, X and Z is .75 | Reliability of Y, X and Z is .85 | Reliability of Y, X and Z is .95



Panel B: Standard error bias of β_3

Reliability of Y, X and Z is .75 | Reliability of Y, X and Z is .85 | Reliability of Y, X and Z is .95

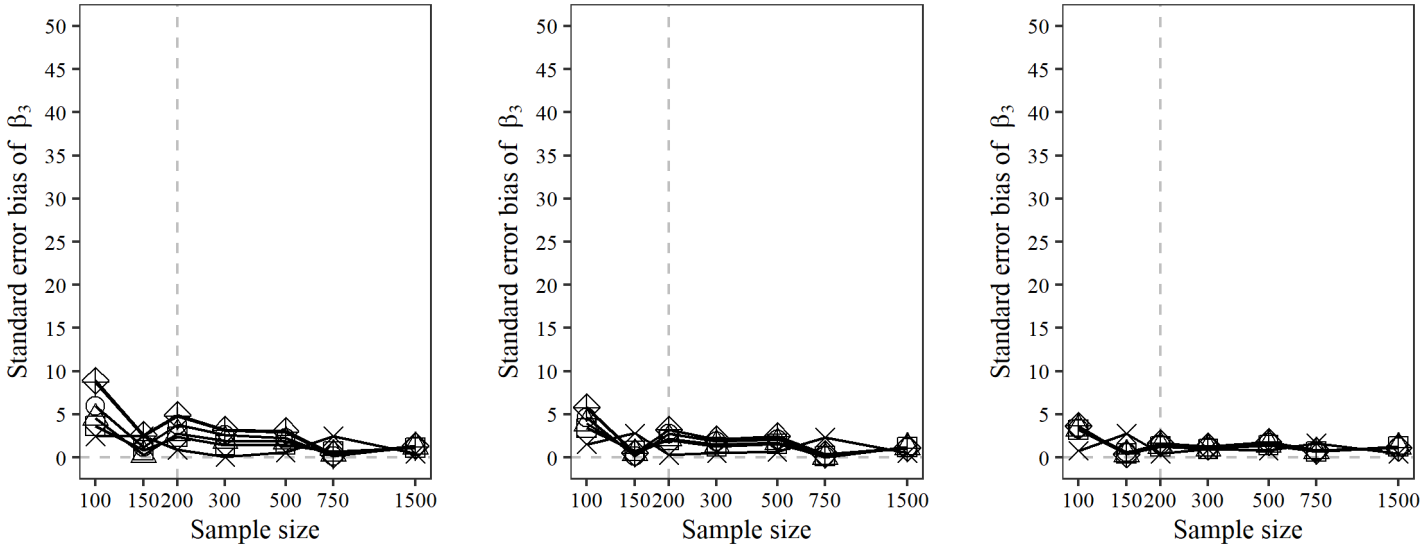
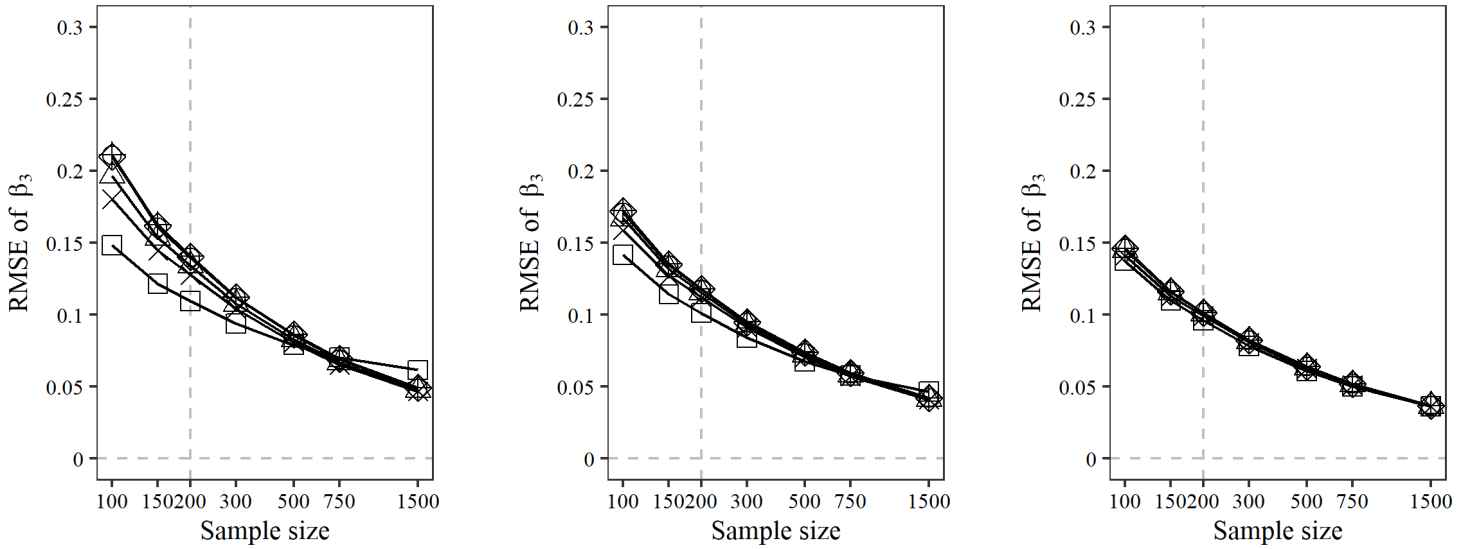


Figure WA6 (CONTINUED)

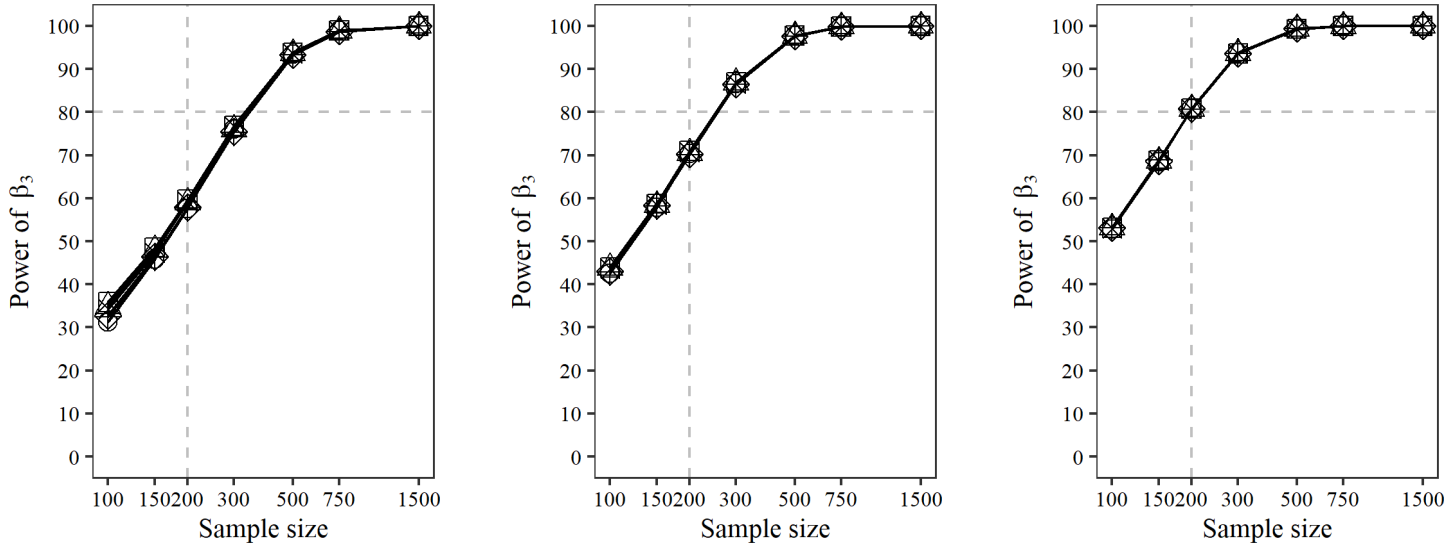
Panel C: RMSE of β_3

Reliability of Y, X and Z is .75 | Reliability of Y, X and Z is .85 | Reliability of Y, X and Z is .95



Panel D: Power of β_3

Reliability of Y, X and Z is .75 | Reliability of Y, X and Z is .85 | Reliability of Y, X and Z is .95



Legend:

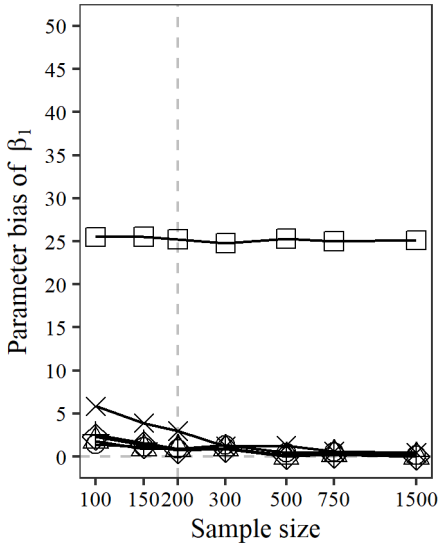
1. Multi-group
2. Means
3. Corrected means
4. Factor scores
5. Product indicators
6. Latent product

Notes: Plots visualize method parameter bias, standard error bias, root mean squared error (RMSE) and power (as defined in Table 2 in the main text) of the moderation effect (β_3) across sample sizes (log scale) and reliabilities of Y, X and Z. Horizontal dashed lines indicate parameter bias, standard error bias and RMSE of zero and power of 80%. Vertical dashed lines indicate a sample size of 200, about the median in the literature review (Table 1 in the main text).

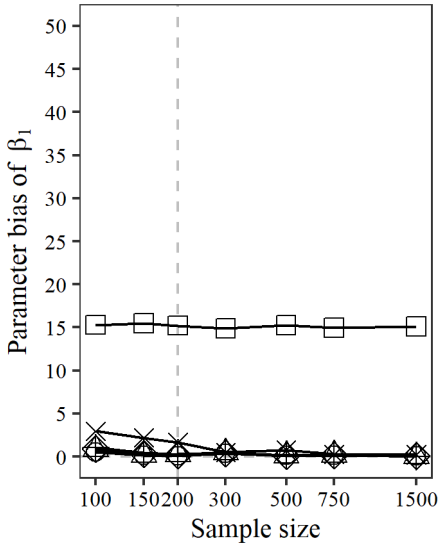
Figure WA7
Study 2b: Performance Criteria for the Main Effect of X (β_1)

Panel A: Parameter bias of β_1

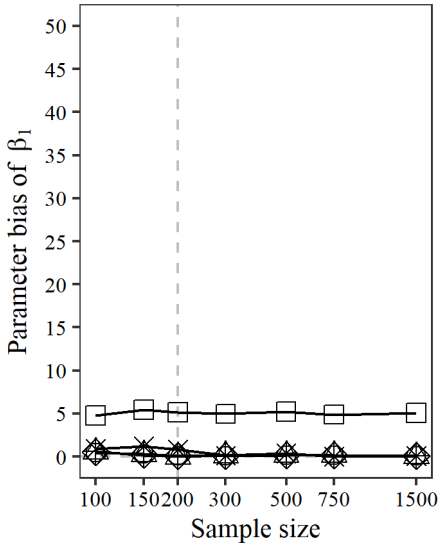
Reliability of Y, X and Z is .75



Reliability of Y, X and Z is .85

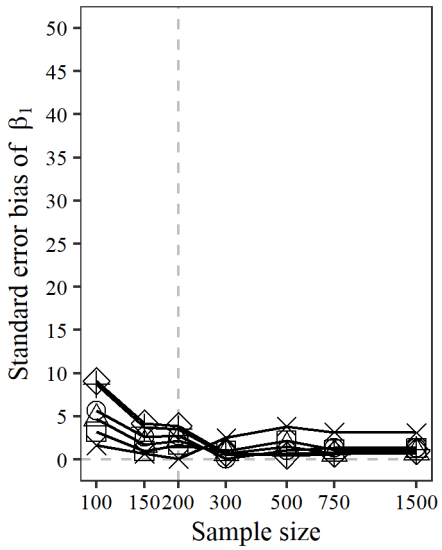


Reliability of Y, X and Z is .95

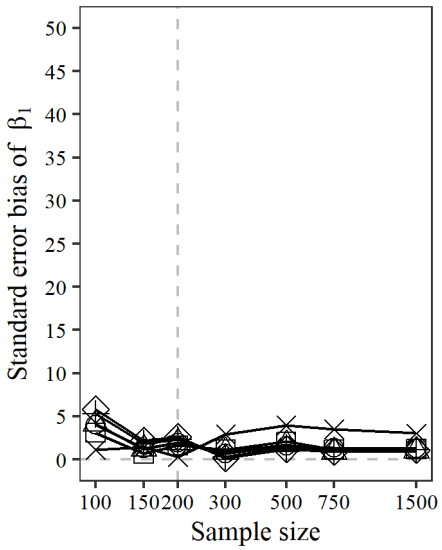


Panel B: Standard error bias of β_1

Reliability of Y, X and Z is .75



Reliability of Y, X and Z is .85



Reliability of Y, X and Z is .95

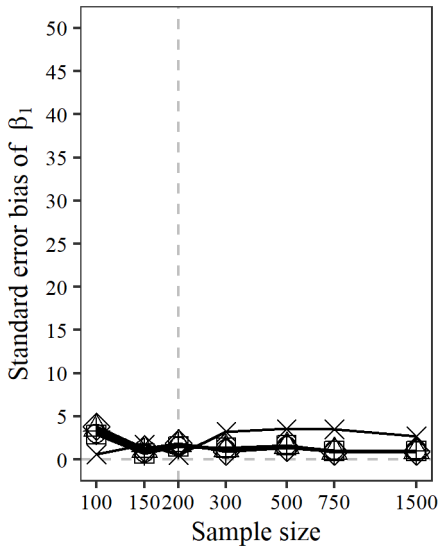
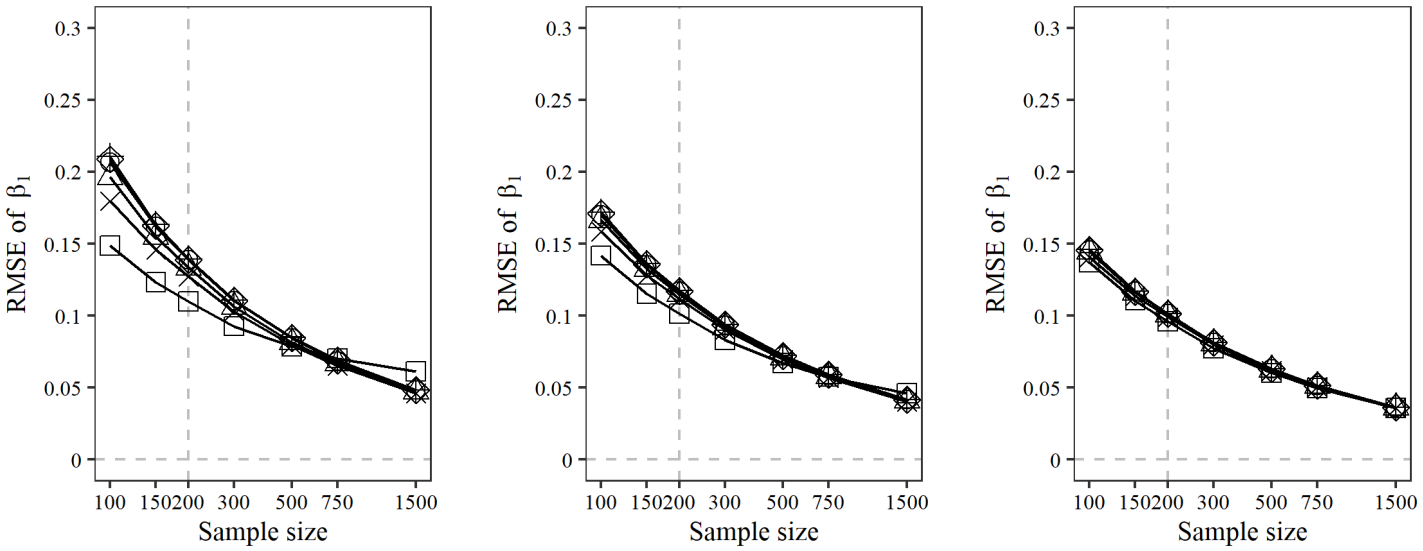


Figure WA7 (CONTINUED)

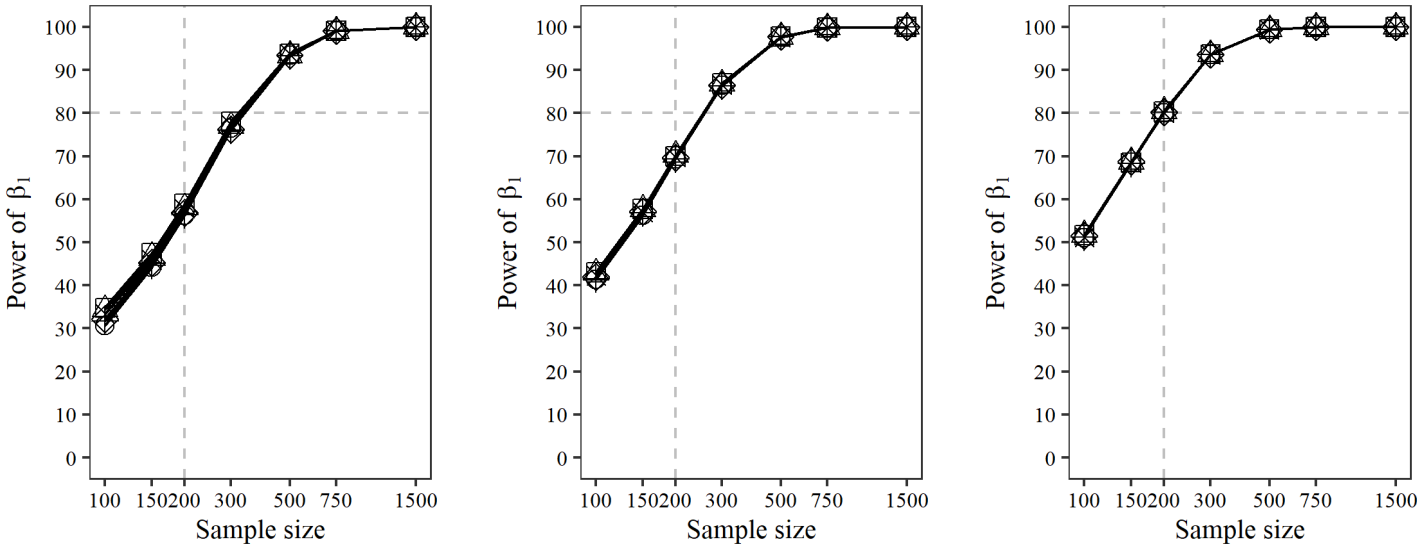
Panel C: RMSE of β_1

Reliability of Y, X and Z is .75 | Reliability of Y, X and Z is .85 | Reliability of Y, X and Z is .95



Panel D: Power of β_1

Reliability of Y, X and Z is .75 | Reliability of Y, X and Z is .85 | Reliability of Y, X and Z is .95



Legend:

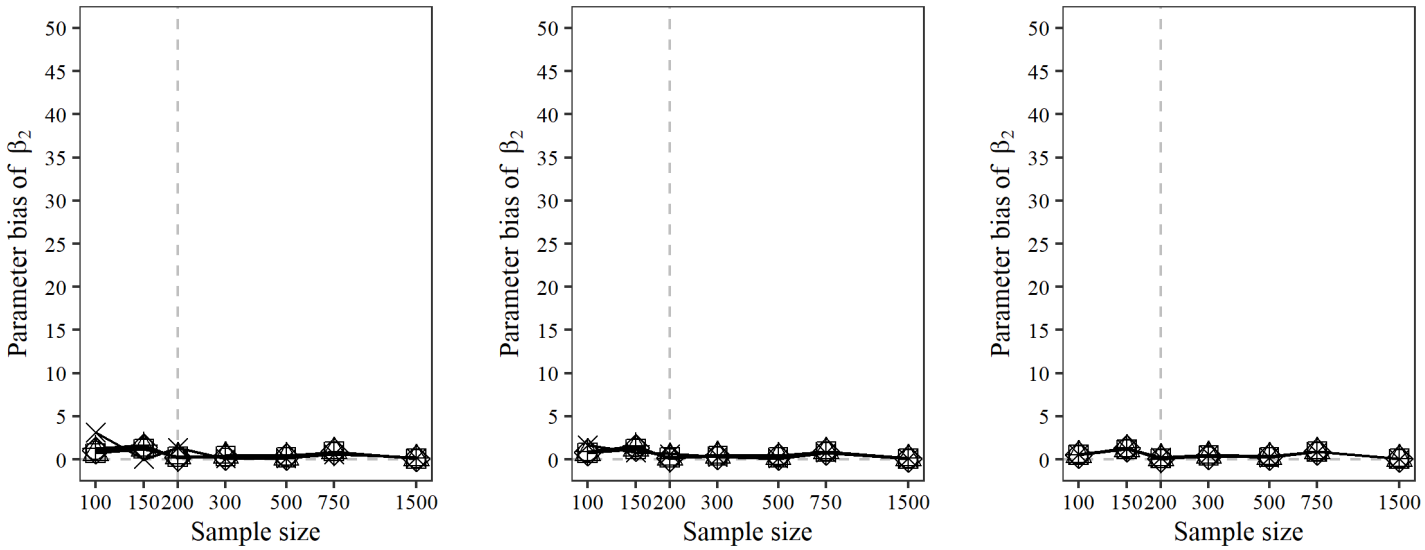
1. Multi-group
2. Means
3. Corrected means
4. Factor scores
5. Product indicators
6. Latent product

Notes: Plots visualize method parameter bias, standard error bias, root mean squared error (RMSE) and power (as defined in Table 2 in the main text) of the main effect of X (β_1) across sample sizes (log scale) and reliabilities of Y, X and Z. Horizontal dashed lines indicate parameter bias, standard error bias and RMSE of zero and power of 80%. Vertical dashed lines indicate a sample size of 200, about the median in the literature review (Table 1 in the main text).

Figure WA8
Study 2b: Performance Criteria for the Main Effect of Z (β_2)

Panel A: Parameter bias of β_2

Reliability of Y, X and Z is .75	Reliability of Y, X and Z is .85	Reliability of Y, X and Z is .95



Panel B: Standard error bias of β_2

Reliability of Y, X and Z is .75	Reliability of Y, X and Z is .85	Reliability of Y, X and Z is .95

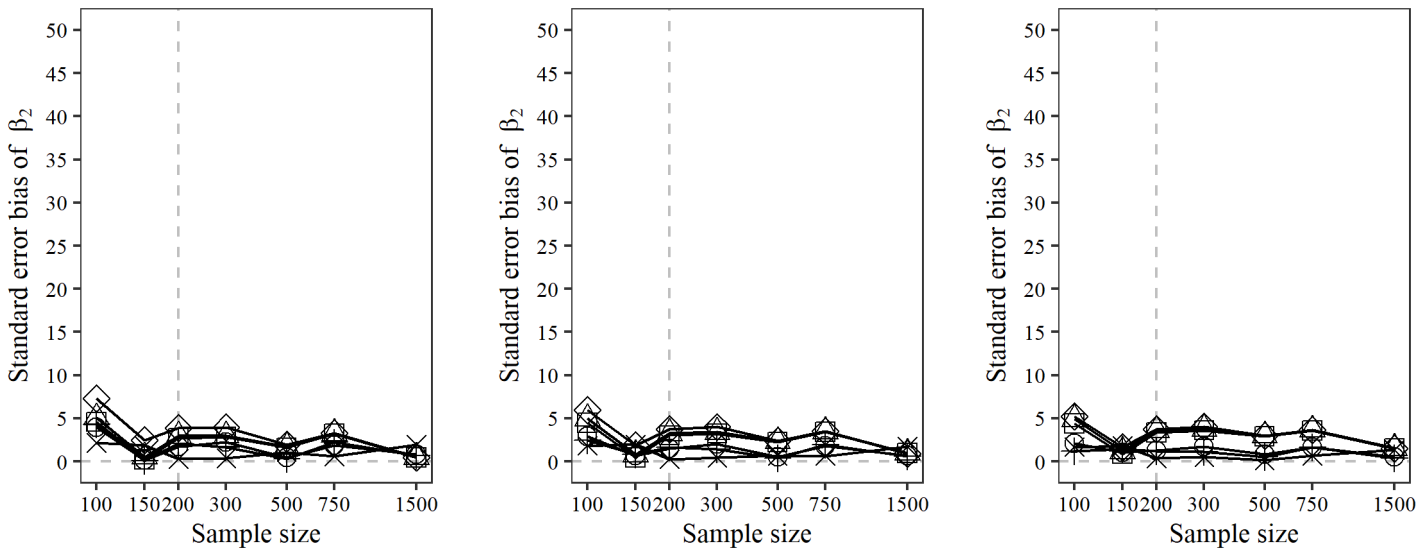
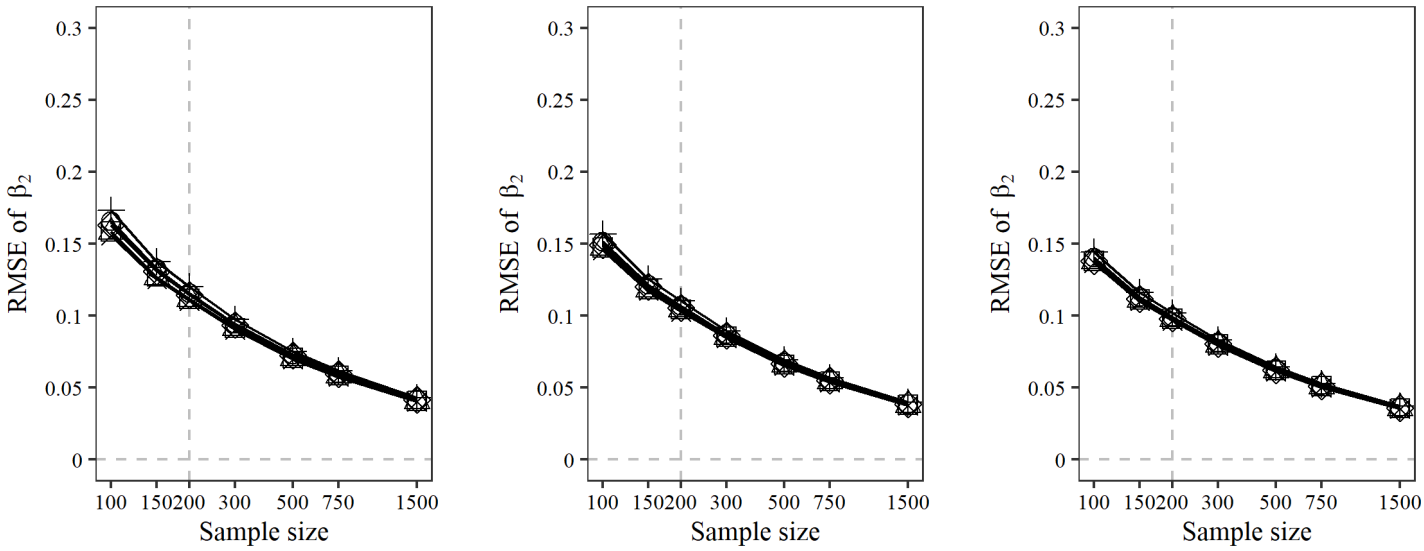


Figure WA8 (CONTINUED)

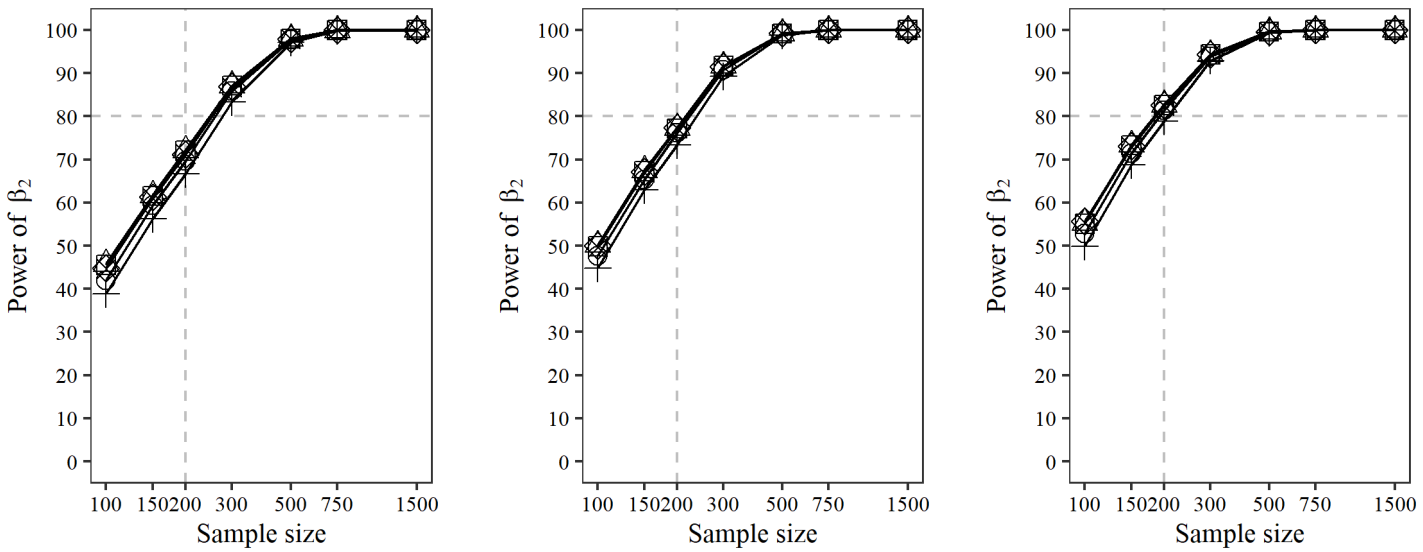
Panel C: RMSE of β_2

Reliability of Y, X and Z is .75 | Reliability of Y, X and Z is .85 | Reliability of Y, X and Z is .95



Panel D: Power of β_2

Reliability of Y, X and Z is .75 | Reliability of Y, X and Z is .85 | Reliability of Y, X and Z is .95



Legend:

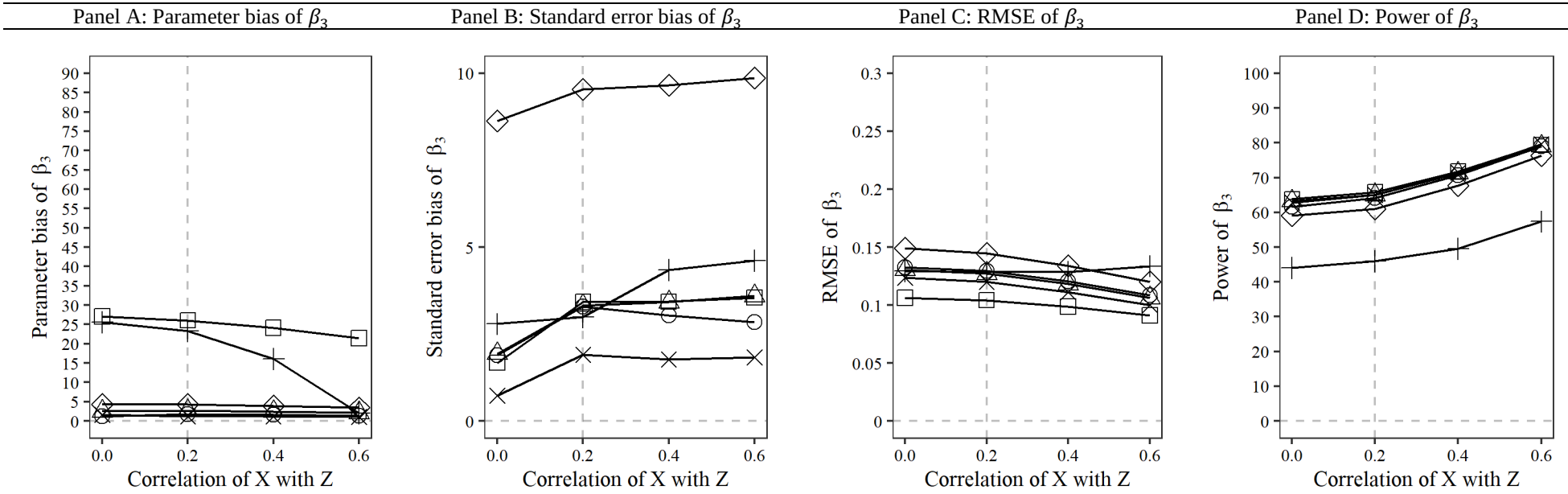
1. Multi-group
2. Means
3. Corrected means
4. Factor scores
5. Product indicators
6. Latent product

Notes: Plots visualize method parameter bias, standard error bias, root mean squared error (RMSE) and power (as defined in Table 2 in the main text) of the main effect of Z (β_2) across sample sizes (log scale) and reliabilities of Y, X and Z. Horizontal dashed lines indicate parameter bias, standard error bias and RMSE of zero and power of 80%. Vertical dashed lines indicate a sample size of 200, about the median in the literature review (Table 1 in the main text).

Web Appendix I: Study 2c

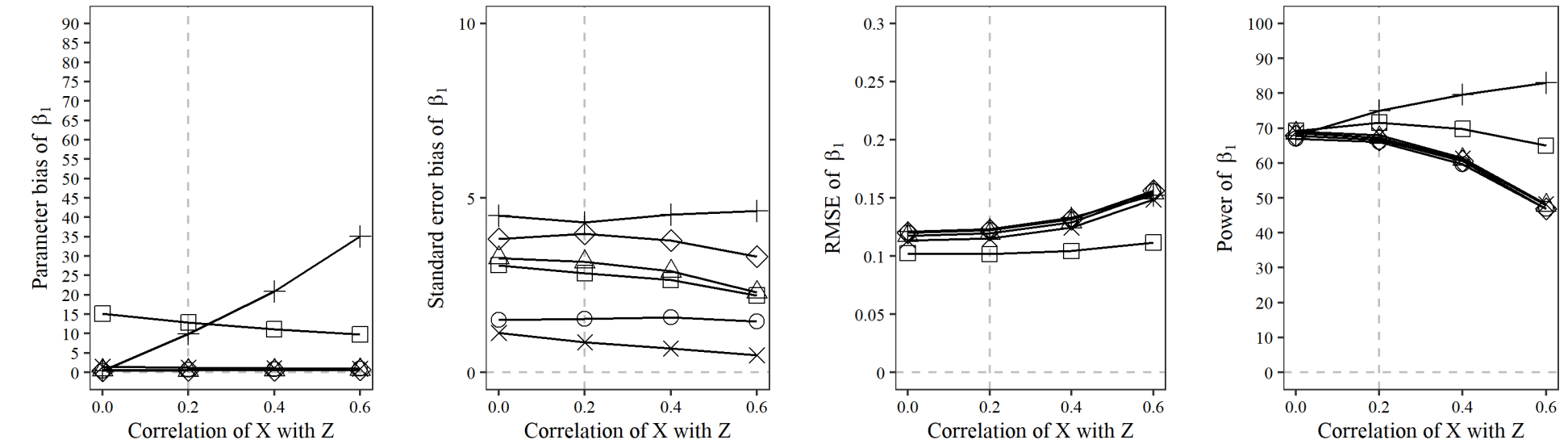
Figures WA9-11 have detailed results.

Figure WA9
Study 2c: Performance Criteria for the Moderation Effect (β_3)



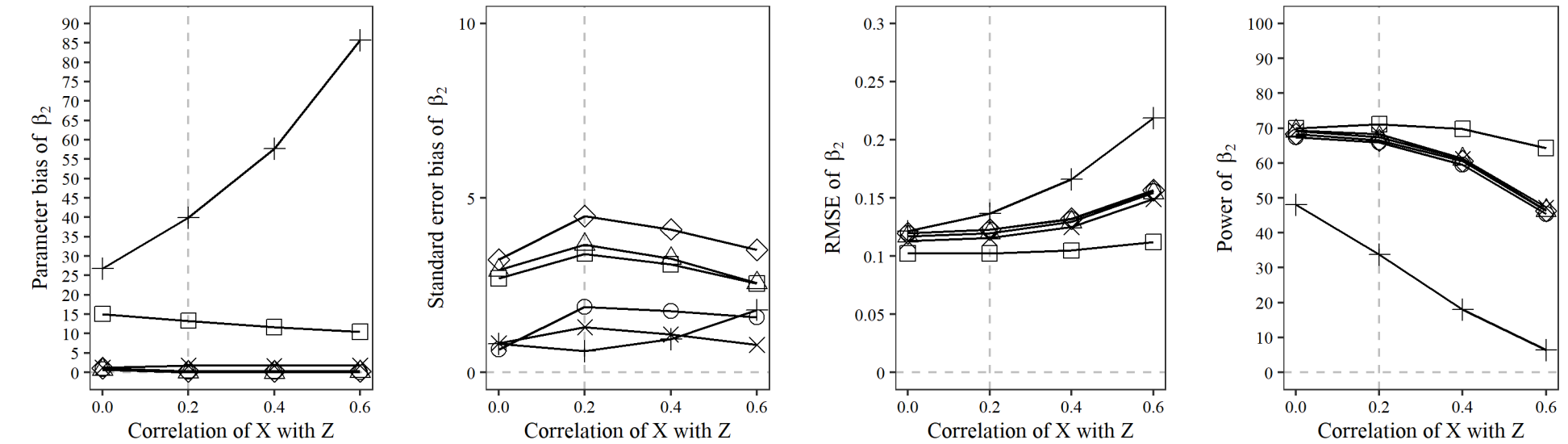
Notes: Plots visualize method parameter bias, standard error bias, root mean squared error (RMSE) and power (as defined in Table 2 in the main text) of the moderation effect (β_3) across sample sizes (log scale). Horizontal dashed lines indicate parameter bias, standard error bias and RMSE of zero and power of 80%. Vertical dashed lines indicate a sample size of 200, about the median in the literature review (Table 1 in the main text).

Figure WA10
Study 2c: Performance Criteria for the Main Effect of X (β_1)



Notes: Plots visualize method parameter bias, standard error bias, root mean squared error (RMSE) and power (as defined in Table 2 in the main text) of the main effect of X (β_1) across sample sizes (log scale). Horizontal dashed lines indicate parameter bias, standard error bias and RMSE of zero and power of 80%. Vertical dashed lines indicate a sample size of 200, about the median in the literature review (Table 1 in the main text).

Figure WA11
Study 2c: Performance Criteria for the Main Effect of Z (β_2)

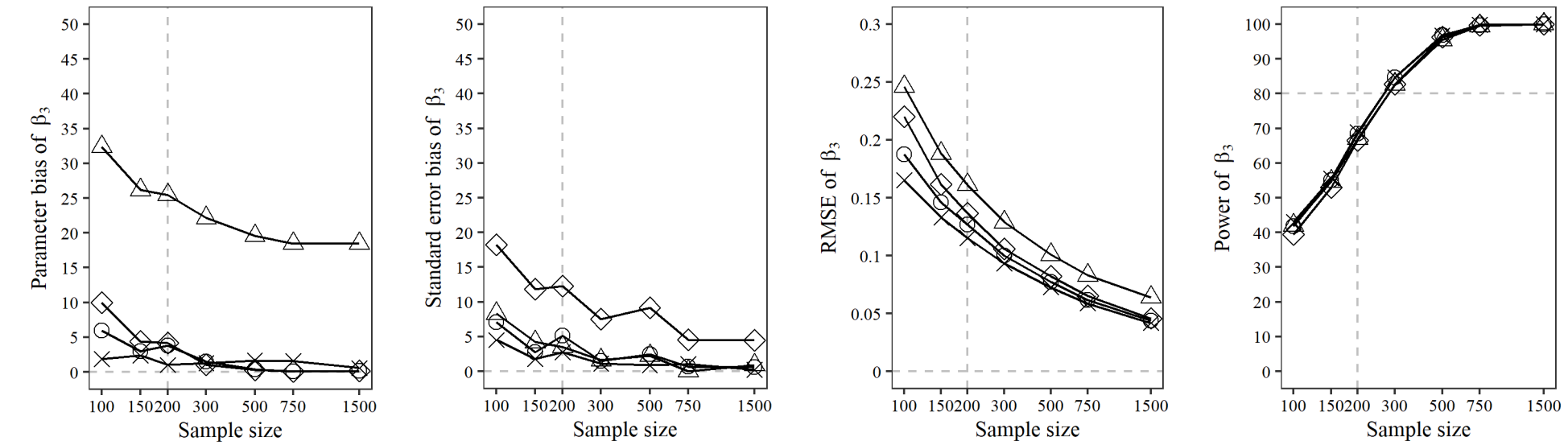


Notes: Plots visualize method parameter bias, standard error bias, root mean squared error (RMSE) and power (as defined in Table 2 in the main text) of the main effect of Z (β_2) across sample sizes (log scale). Horizontal dashed lines indicate parameter bias, standard error bias and RMSE of zero and power of 80%. Vertical dashed lines indicate a sample size of 200, about the median in the literature review (Table 1 in the main text).

Web Appendix J: Study 3

Figure WA12 has detailed results. OSF has the performance criteria for the main effects.

Figure WA12
Study 3: Performance Criteria for the Moderation Effect (β_3)



Notes: Plots visualize method parameter bias, standard error bias, root mean squared error (RMSE) and power (as defined in Table 2 in the main text) of the moderation effect (β_3) across sample sizes (log scale). Horizontal dashed lines indicate parameter bias, standard error bias and RMSE of zero and power of 80%. Vertical dashed lines indicate a sample size of 200, about the median in the literature review (Table 1 in the main text).

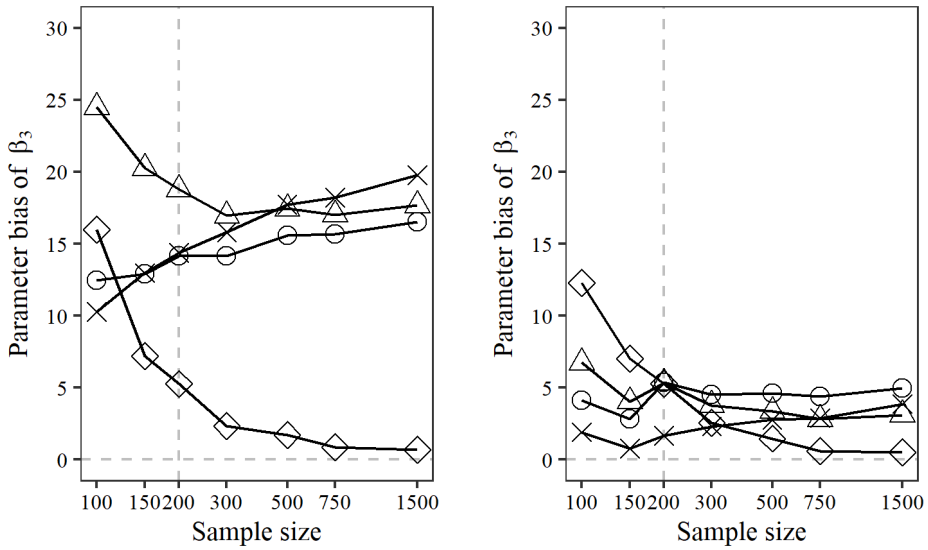
Web Appendix K: Study 4a

Figure WA13 has detailed results. OSF has the performance criteria for the main effects.

Figure WA13
Study 4a: Performance Criteria for the Moderation Effect (β_3)

Panel A: Parameter bias of β_3

Skewness 3, excess kurtosis 10 (X and Z) | Skewness 1, excess kurtosis 2 (X and Z)



Panel B: Standard error bias of β_3

Skewness 3, excess kurtosis 10 (X and Z) | Skewness 1, excess kurtosis 2 (X and Z)

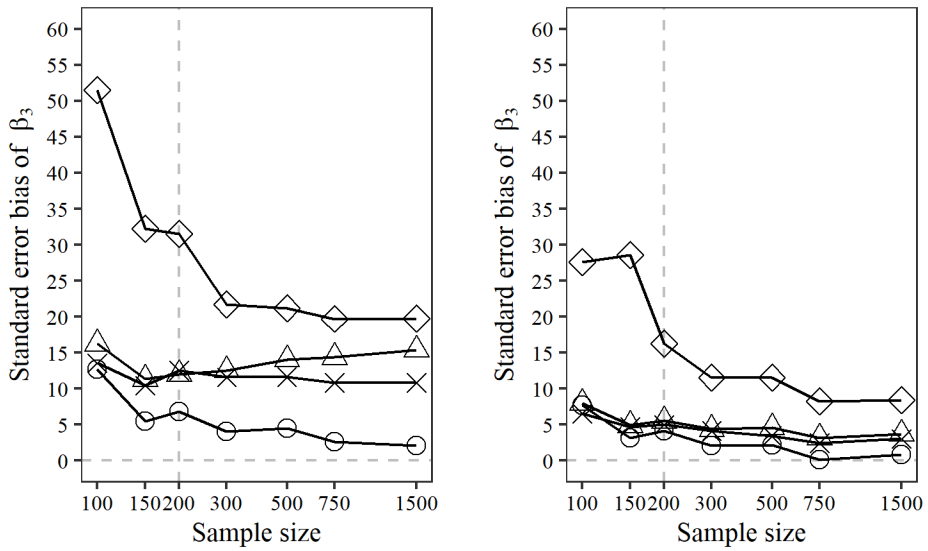
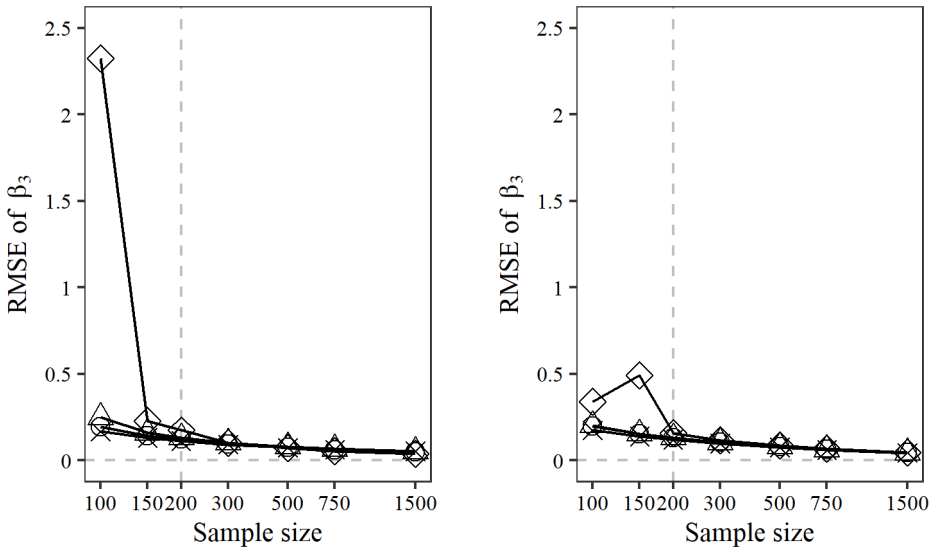


Figure WA13 (CONTINUED)

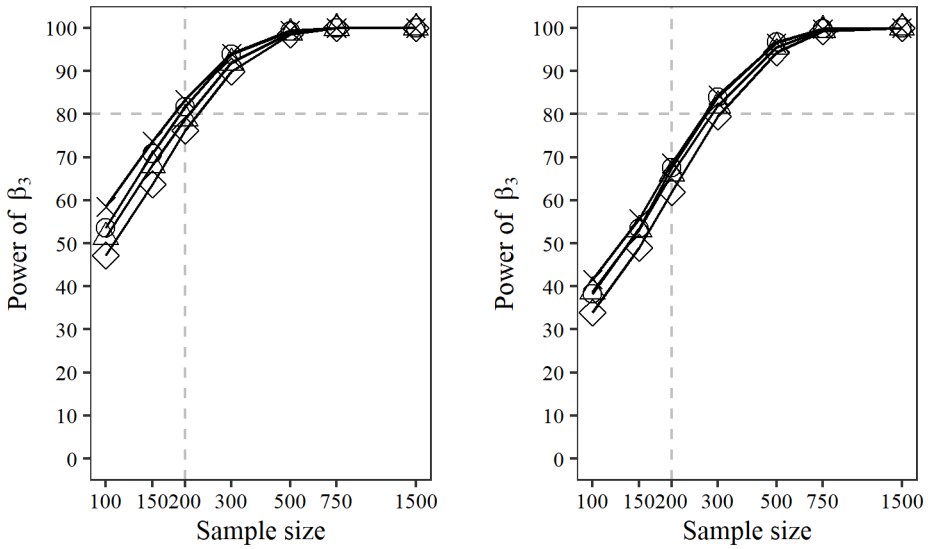
Panel C: RMSE of β_3

Skewness 3, excess kurtosis 10 (X and Z) Skewness 1, excess kurtosis 2 (X and Z)



Panel D: Power of β_3

Skewness 3, excess kurtosis 10 (X and Z) Skewness 1, excess kurtosis 2 (X and Z)



Legend:

△ 3. Corrected means × 4. Factor scores ◇ 5. Product indicators ○ 6. Latent product

Notes: Plots visualize method parameter bias, standard error bias, root mean squared error (RMSE) and power (as defined in Table 2 in the main text) of the moderation effect (β_3) across sample sizes (log scale) and levels of non-normality of X and Z. Horizontal dashed lines indicate parameter bias, standard error bias and RMSE of zero and power of 80%. Vertical dashed lines indicate a sample size of 200, about the median in the literature review (Table 1 in the main text).

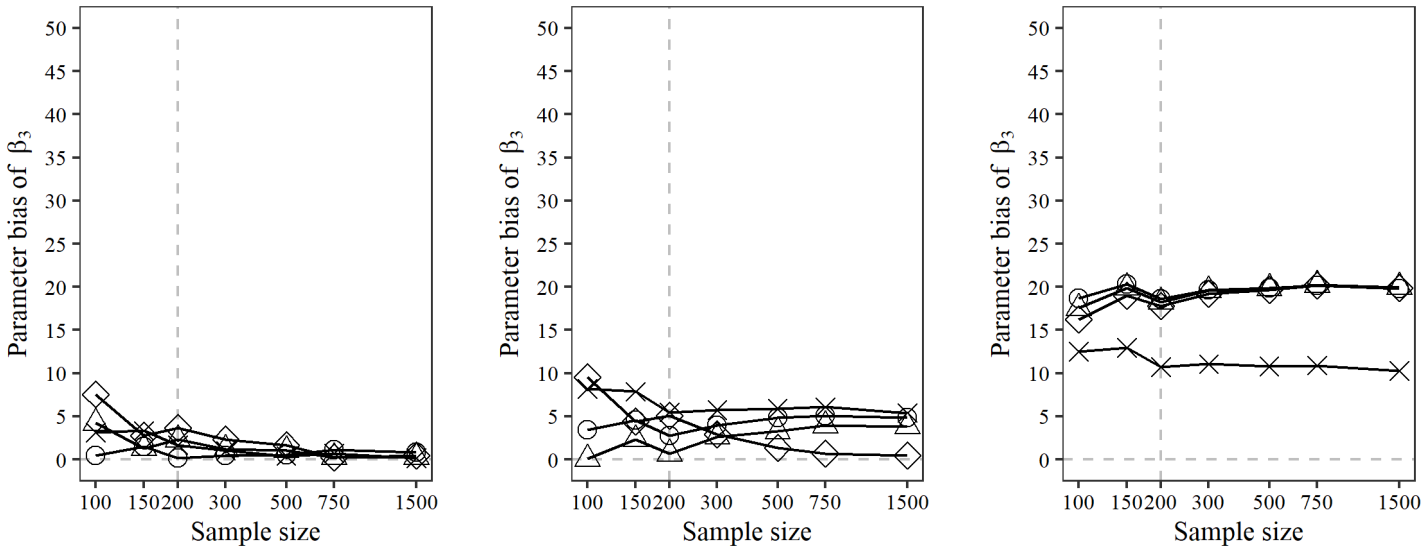
Web Appendix L: Study 4b

Figures WA14-15 have detailed results. OSF has the performance criteria for the main effect of Z.

Figure WA14
Study 4b: Performance Criteria for the Moderation Effect (β_3)

Panel A: Parameter bias of β_3

Correlated measurement errors of x with y	Correlated measurement errors of x with z	Correlated measurement errors of x with x



Panel B: Standard error bias of β_3

Correlated measurement errors of x with y	Correlated measurement errors of x with z	Correlated measurement errors of x with x

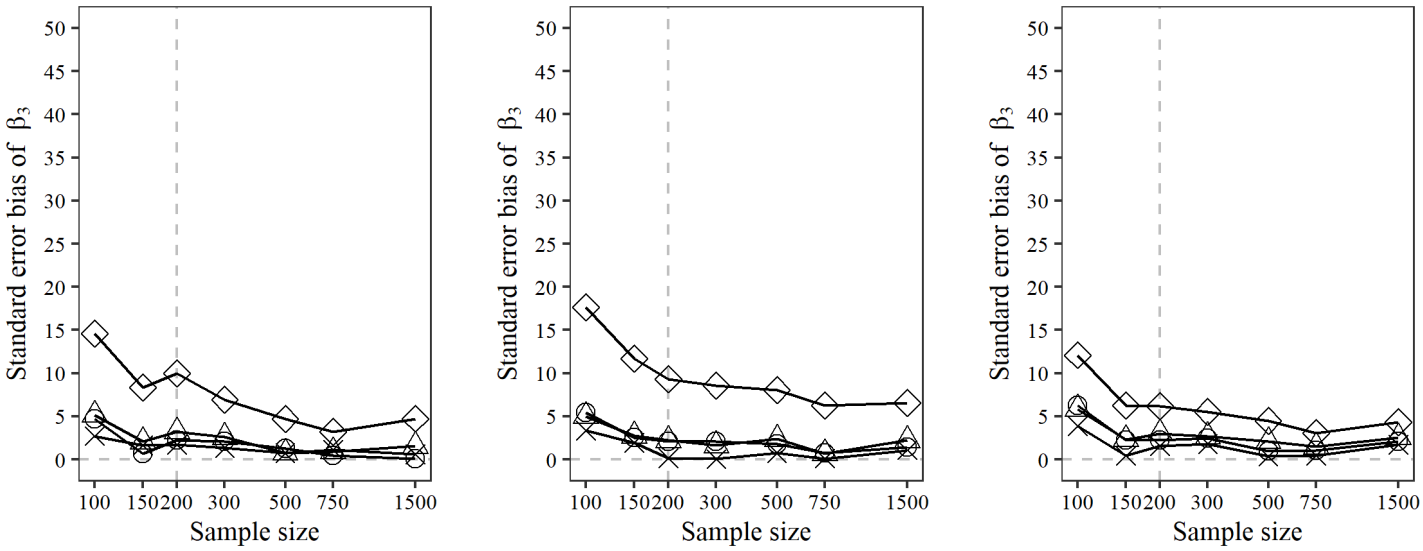
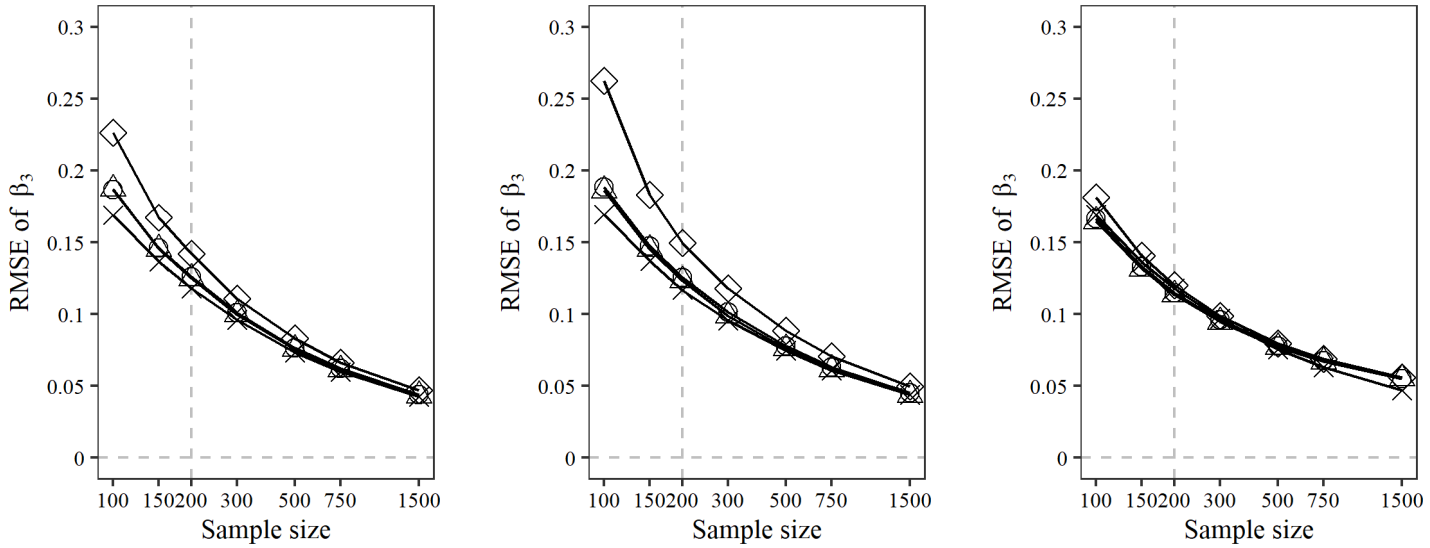


Figure WA14 (CONTINUED)

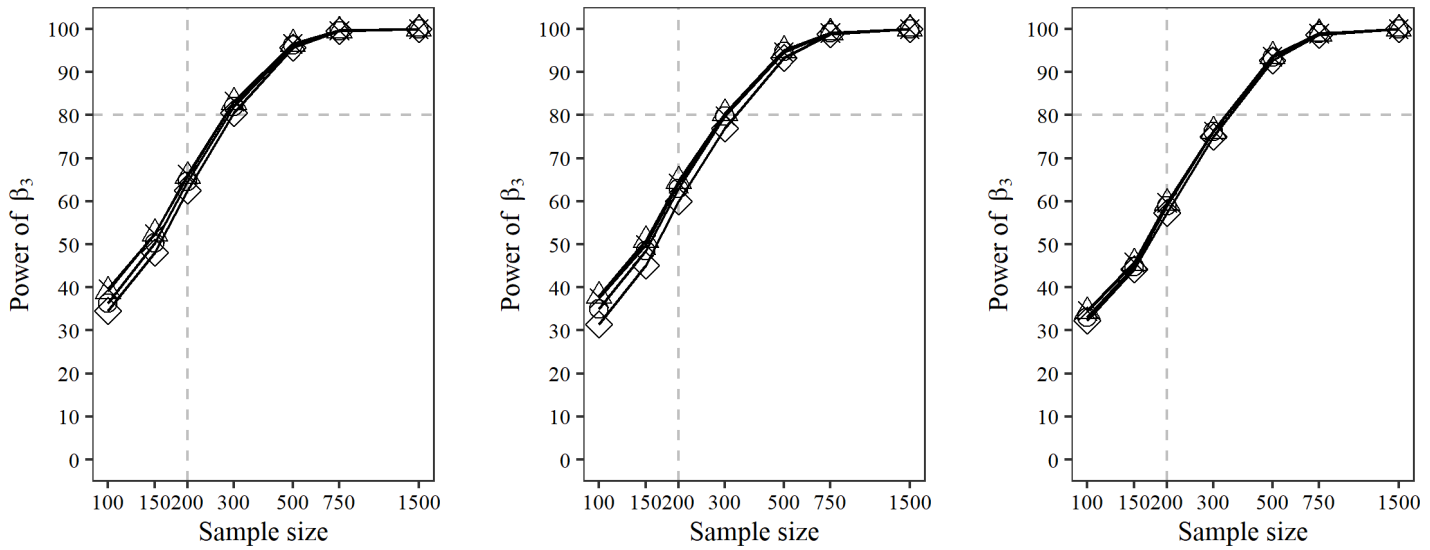
Panel C: RMSE of β_3

Correlated measurement errors of x with y	Correlated measurement errors of x with z	Correlated measurement errors of x with x



Panel D: Power of β_3

Correlated measurement errors of x with y	Correlated measurement errors of x with z	Correlated measurement errors of x with x



Legend:

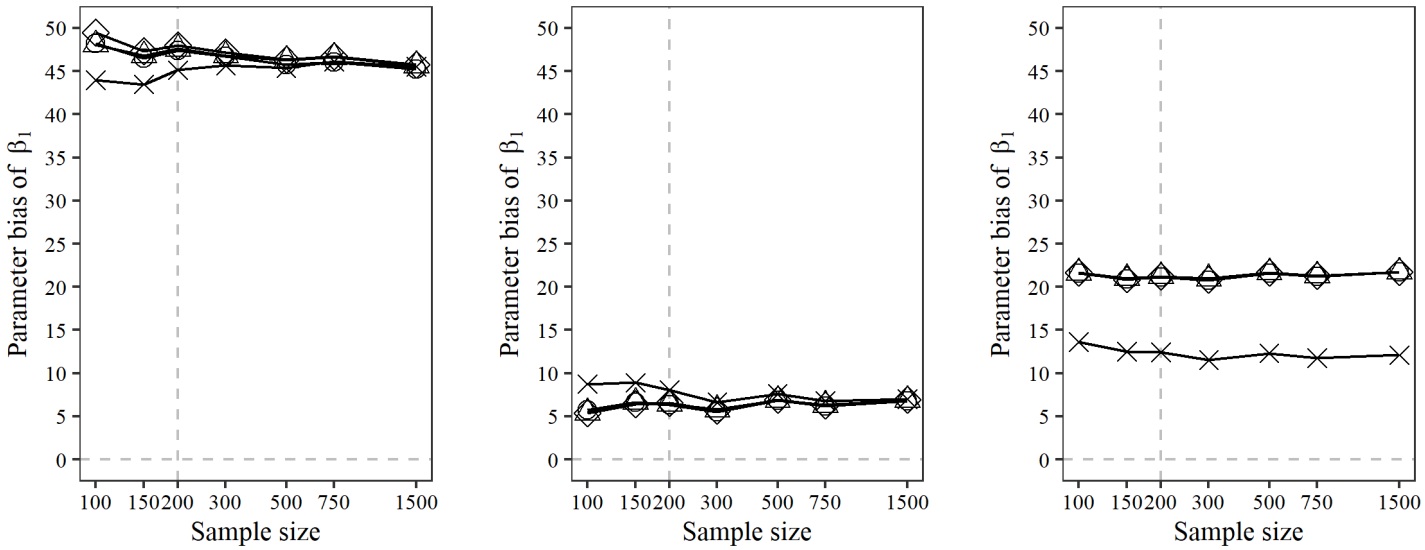
△ 3. Corrected means × 4. Factor scores ◇ 5. Product indicators ○ 6. Latent product

Notes: Plots visualize method parameter bias, standard error bias, root mean squared error (RMSE) and power (as defined in Table 2 in the main text) of the moderation effect (β_3) across sample sizes (log scale) and types of correlated measurement errors. Horizontal dashed lines indicate parameter bias, standard error bias and RMSE of zero and power of 80%. Vertical dashed lines indicate a sample size of 200, about the median in the literature review (Table 1 in the main text).

Figure WA15
Study 4b: Performance Criteria for the Main Effect of X (β_1)

Panel A: Parameter bias of β_1

Correlated measurement errors of x with y	Correlated measurement errors of x with z	Correlated measurement errors of x with x



Panel B: Standard error bias of β_1

Correlated measurement errors of x with y	Correlated measurement errors of x with z	Correlated measurement errors of x with x

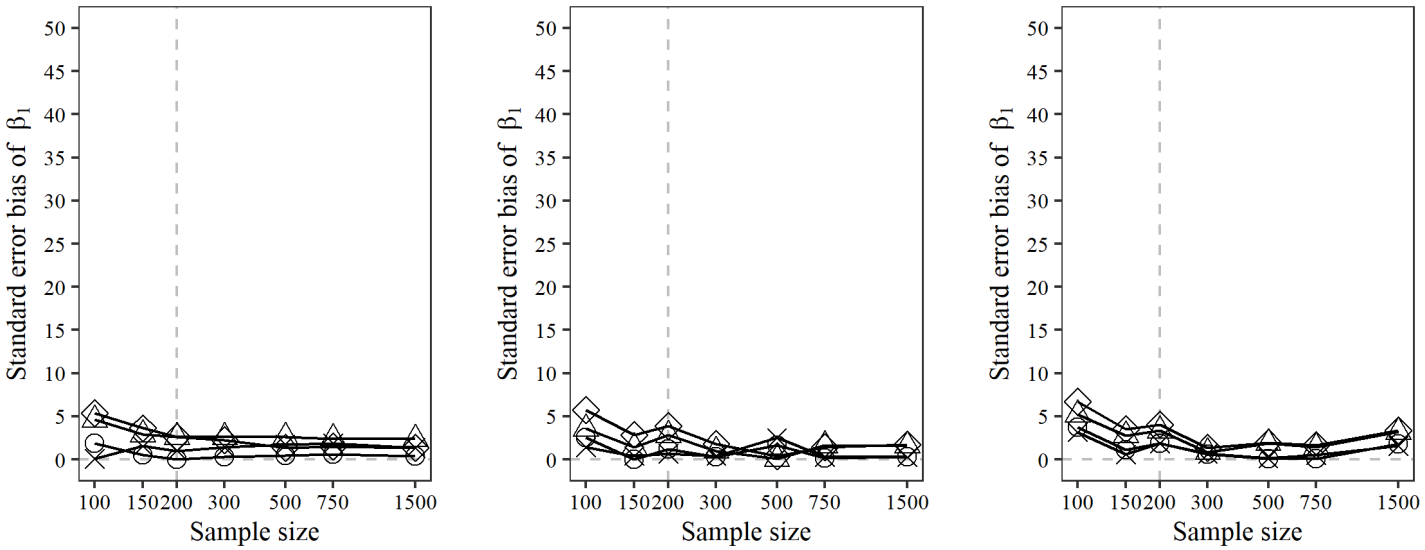
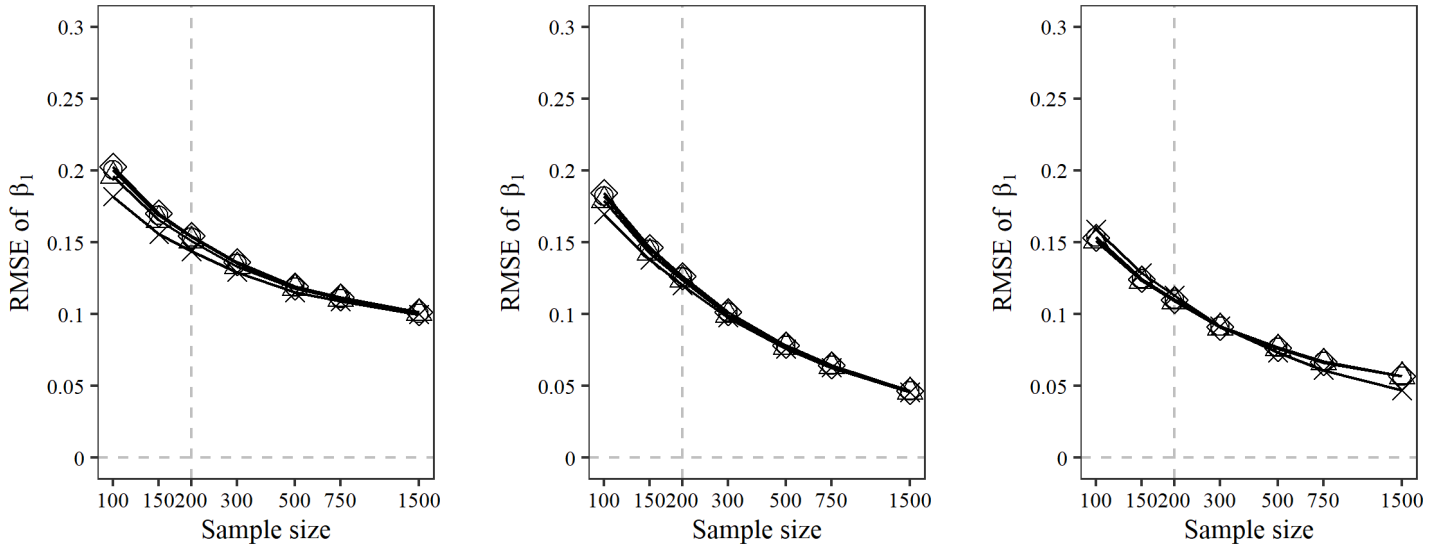


Figure WA15 (CONTINUED)

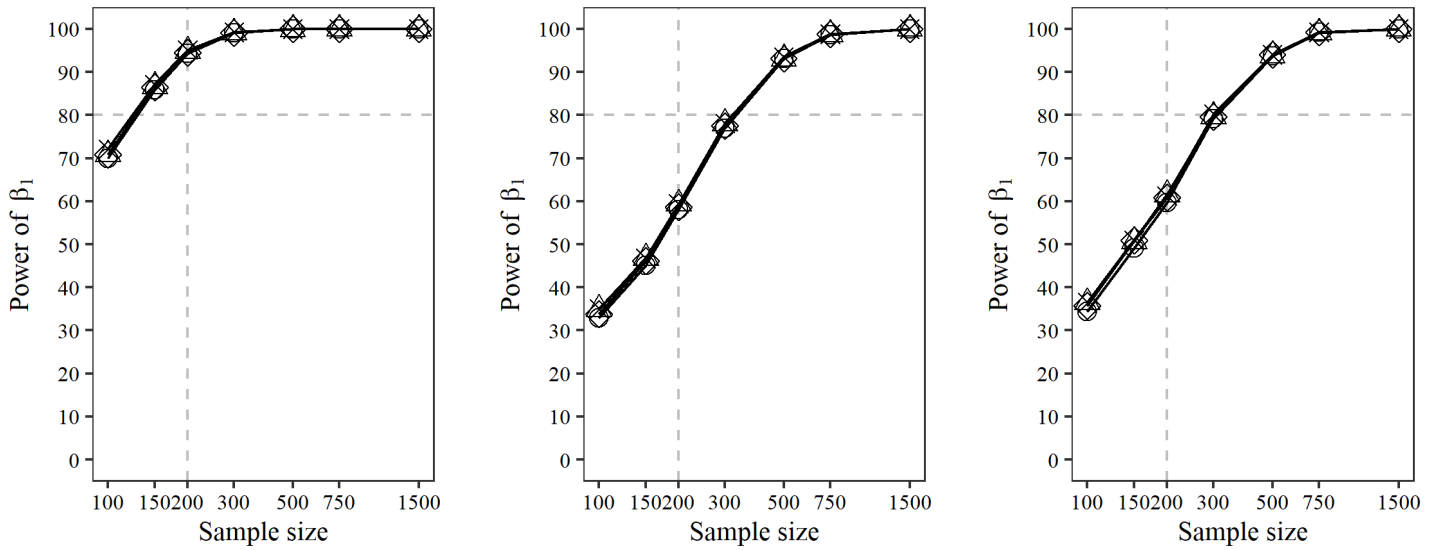
Panel C: RMSE of β_3

Correlated measurement errors of x with y	Correlated measurement errors of x with z	Correlated measurement errors of x with x



Panel D: Power of β_1

Correlated measurement errors between x and y	Correlated measurement errors between x and z	Correlated measurement errors between x and x



Legend:

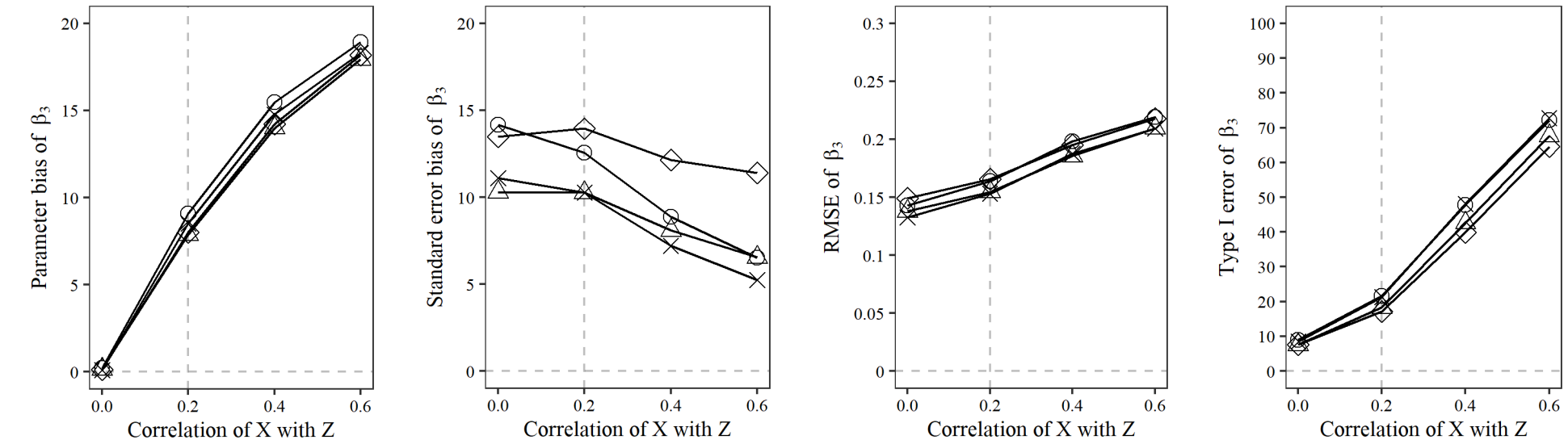
△ 3. Corrected means × 4. Factor scores ◇ 5. Product indicators ○ 6. Latent product

Notes: Plots visualize method parameter bias, standard error bias, root mean squared error (RMSE) and power (as defined in Table 2 in the main text) of the main effect of X (β_1) across sample sizes (log scale) and measurement error correlations. Horizontal dashed lines indicate parameter bias, standard error bias and RMSE of zero and power of 80%. Vertical dashed lines indicate a sample size of 200, about the median in the literature review (Table 1 in the main text).

Web Appendix M: Study 4c

Figure WA16 has detailed results. OSF has the performance criteria for the main effects.

Figure WA16
Study 4c: Performance Criteria for the Moderation Effect (β_3)



Notes: Plots visualize method parameter bias, standard error bias, root mean squared error (RMSE) and type I error (as defined in Table 2 in the main text) of the moderation effect (β_3) across the correlation of X with Z. Horizontal dashed lines indicate parameter bias, standard error bias, RMSE and type I error of zero. Vertical dashed lines indicate a correlation of X with Z of .20, about the mean from the literature review (Table 1 in the main text).

Web Appendix N: Follow-up Study 1

The objective is to investigate differences in estimation time between the latent product method implemented in Mplus (Muthén and Muthén 2019), and the R implementation in package nlsem (Umbach et al. 2017).

Method

The design has two groups (Latent product method implementation: Mplus or R-nlsem). We generate 5,000 datasets with a sample size of 200, reliabilities of Y, X and Z of .85, a correlation between X and Y of .20 and .20 regression weights for main and moderation effects. Estimation settings and criteria for model convergence are held equal such as the use of 16 mixture components and a required change in log-likelihood smaller than .001 for convergence, with a maximum of 500 iterations. We use a personal computer with an Intel Core i7-4790 CPU running at 3.6 GHz and 32GB RAM for estimation. For each replication, we save the estimation time in seconds.

Results

Table WA4 has the results. Panel A shows large differences in estimation time for the latent product method implementations in Mplus and R-nlsem. Estimation for the latent product method in Mplus took an average of two to three seconds (M = 2.67, Mdn = 1.70, SD = 8.27) while R-nlsem estimation took close to a minute (M = 56.51, Mdn = 55.03, SD = 14.30). Thus, Mplus estimates the parameters much quicker than R-nlsem does.

To investigate whether the much shorter estimation timing also harms parameter recovery, we also look at the focal performance criteria. Note that we keep the number of mixture components (i.e., estimation precision) constant between the implementations such that it could not account for any differences. Although we find limited parameter bias for both Mplus and R-nlsem implementations (< 1.5% for all βs), the R-nlsem implementation

has standard error bias, about 10% for β_1 and β_3. The Mplus implementation has much less bias, a maximum about 2%. This also results in lower RMSE for R-nlsem and a higher power compared to Mplus.

In sum, the R-nlsem implementation is not only slower than Mplus, it also performs substantively worse in terms of standard error bias. Thus, the Mplus implementation is preferred over nlsem. Other advantages are the use of the latent product method with single-indicators (Hsiao et al. 2021), categorical moderation variables (Muthén and Muthén 2019) and the availability of Bayesian estimation for the latent product method (Asparouhov and Muthén 2021). A feature of nlsem is the availability of the quasi-maximum likelihood (QML) estimation of the latent product method (Klein and Muthén 2007), which has a less computationally intensive estimation algorithm but is less precise than the expectation maximization estimation focused on here (Kelava et al. 2011, p. 476).

Table WA4
The Latent Product Method Implemented in Mplus Outperforms R-nlsem

Panel A: Estimation timing (in seconds)			
Implementation	M	Mdn	SD
Mplus	2.667	1.704	8.271
R-nlsem	56.508	55.031	14.303

Panel B: Performance criteria				
Effect and Implementation	Par. Bias	SE Bias	RMSE	Power
Main effect of X on Y (β_1)				
Mplus	1.226	1.259	.122	66.544
R-nlsem	1.322	10.217	.116	73.132
Main effect of X on Y (β_2)				
Mplus	.343	1.176	.122	66.359
R-nlsem	.332	2.741	.121	67.549
Moderation effect of XZ on Y (β_3)				
Mplus	.05	2.027	.128	62.787
R-nlsem	.014	10.212	.122	69.602

Notes: M is the mean, Mdn is the median and SD is the standard deviation of estimation timing in seconds. Par. Bias refers to parameter bias, SE Bias to standard error bias, RMSE is the root mean squared error of β and Power refers to the statistical power of β, as defined in Table 2 in the main text.

Web Appendix O: Follow-up Study 2

The latent product method uses a mixture distribution to approximate the non-normal indicator distribution due to the interaction. The number of normal distributions, or mixture components, trade off precision with computational intensiveness. We use 16 mixture components for the latent product method, following recommendations by Klein and Moosbrugger (2000, p. 465), for all our studies. Yet, the question remains whether the results are sensitive to the number of mixture components.

Method

The design has 5 groups (Number of mixture components: 2, 9, 16, 23 or 30). We generate 5,000 datasets with a sample size of 200, reliabilities of Y, X and Z of .85, a correlation between X and Y of .20 and .20 regression weights for main and moderation effects. We then use between 2 and 30 (in steps of 7) mixture components for estimation with the latent product method (the default in Mplus is 15 mixture components (Muthén and Muthén 2019) and nlsem uses 16 mixture components by default (Umbach et al. 2017)).

Results

Table WA5 has the results. Using two mixture components, the latent product has a moderate parameter bias (e.g., about 11% for β_3) and standard error bias (e.g., 9% for SE[β_3]). This bias decreases when increasing the number of mixture components. Overall, and across the performance criteria, differences between the estimates from estimation with nine or more mixture components are very small. Importantly, the values of the performance criteria are virtually identical for the results with 16 (the number of mixture components used in our studies) and more components, which is encouraging. In sum, the results of the simulations, that use 16 mixture components, are unlikely to change if additional mixture components are used.

Table WA5
The Results are Robust to the Number of Mixture Components

	Performance criteria			
Number of mixture components	Par. Bias	SE Bias	RMSE	Power
2 mixture components				
Main effect of X on Y (β_1)	3.85	5.48	.09	67.1
Main effect of Z on Y (β_2)	6.65	6.96	.09	67.8
Moderation effect of XZ on Y (β_3)	11.1	8.98	.1	66.7
9 mixture components				
Main effect of X on Y (β_1)	.5	.72	.08	66.3
Main effect of Z on Y (β_2)	.55	1.55	.08	66.4
Moderation effect of XZ on Y (β_3)	3.3	1.97	.09	63.1
16 mixture components				
Main effect of X on Y (β_1)	.45	.72	.08	66.3
Main effect of Z on Y (β_2)	.55	1.55	.08	66.4
Moderation effect of XZ on Y (β_3)	3.3	1.97	.09	63.1
23 mixture components				
Main effect of X on Y (β_1)	.45	.72	.08	66.3
Main effect of Z on Y (β_2)	.55	1.55	.08	66.4
Moderation effect of XZ on Y (β_3)	3.25	1.97	.09	63.1
30 mixture components				
Main effect of X on Y (β_1)	.45	.72	.08	66.3
Main effect of Z on Y (β_2)	.55	1.55	.08	66.4
Moderation effect of XZ on Y (β_3)	3.25	1.97	.09	63.1

Note: Par. Bias refers to parameter bias, SE Bias to standard error bias, RMSE is the root mean squared error of β and Power refers to the statistical power of β, as defined in Table 2 in the main text.

Web Appendix P: Follow-up Study 3

This study explores the sensitivity of the results across factor score estimation methods.

Method

The design is: 4 (Factor scores method: Bartlett-Bartlett 2-CFA, Regression-Regression 2-CFA, Bartlett-Regression 2-CFA, Bartlett-Regression 2-EFA) × 7 (Sample size). It focuses on four factor scores estimation methods. The first method (Bartlett-Bartlett 2-CFA) estimates Bartlett scores for Y and X and Z. Similarly, the second method (Regression-Regression 2-CFA) estimates regression factor scores for Y, X and Z. The third method (Bartlett-Regression 2-CFA) uses the recommended Bartlett scores for Y and regression scores for X and Z (Devlieger et al. 2016; Skrondal and Laake 2001). These methods use a 1-CFA for Y and a 2-CFA (without cross-loadings) for X and Z that accounts for the correlation between X and Z. Finally, we explore Bartlett scores for Y and regression scores for X and Z but taken from an unrotated exploratory factor analysis estimated with maximum likelihood (like the CFA's) that fixes the number of factors to two while including cross-loadings but not accounting for the correlation between X and Z (Bartlett-Regression 2-EFA). For consistency, all methods use a path analysis of factor scores, specified as Equation (1) in the main text, to estimate the moderation and main effects. The sample sizes are 100, 150, 200, 300, 500, 750, 1,500 (as in Study 1). The reliability of Y, X and Z is fixed to .85 and the correlation between X and Z is .20, about the means in the literature review (Table 1 in the main text).

Results

Figure WA17 summarizes the results. Panel A has the parameter bias for the moderation and main effects. It shows that all factor scores methods under investigation are biased except for the recommended Bartlett-Regression 2-CFA method (Devlieger et al. 2016; Skrondal and Laake 2001). The bias of the moderation effect is about 27% for Bartlett-Bartlett 2-CFA,

15% for Regression-Regression 2-CFA, < 1% for Bartlett-Regression 2-CFA and 86% for Bartlett-Regression 2-EFA at a sample size of 1,500. The larger bias of the Bartlett-Bartlett 2-CFA is that it accounts for measurement error in Y only and not in XZ, while Regression-Regression accounts for unreliability in XZ only. In this study, the reliability of XZ is smaller than that of Y (cf. Equation (6) in the main text) hence the bias is larger. Inspection of the results reveals that the bias of the Bartlett-Regression 2-EFA method is upward for the main effect of X and downward for the main effect of Z and the moderation effect. This is likely due to the EFA accounting for cross-loadings instead of the correlation between the factors, such that the first factor (for X) accounts for part of the variance of Z and thus overestimating the effect of X and underestimating the remaining effects of Z and XZ on Y.

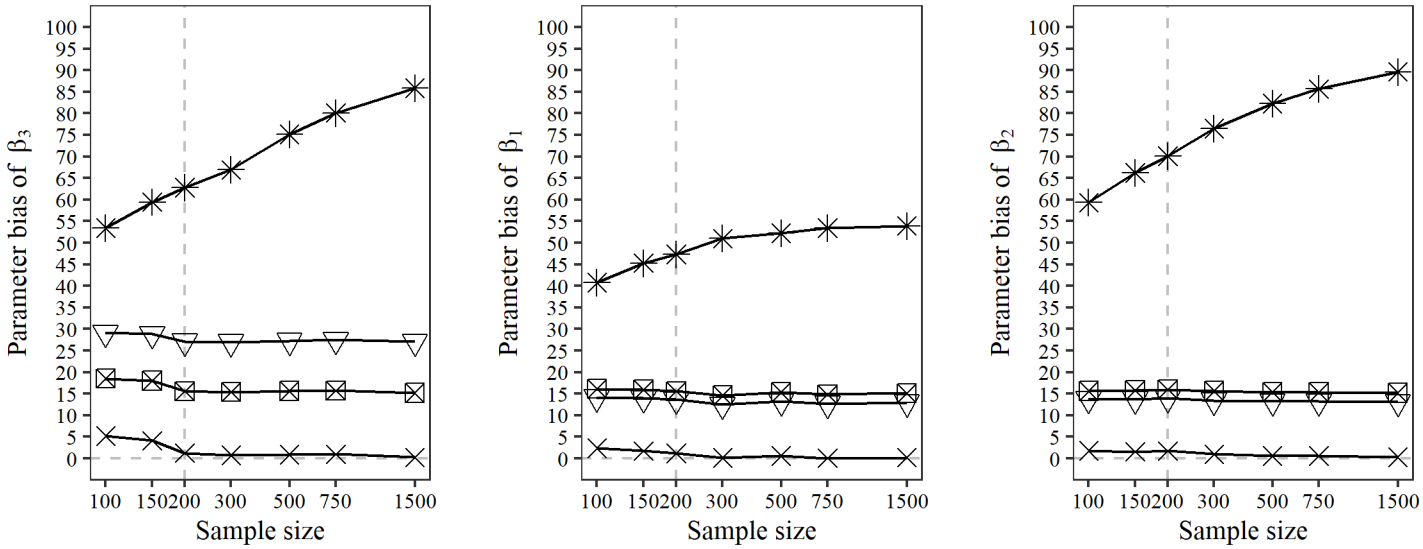
Panel B plots standard error bias. It shows that the standard error bias of Bartlett-Regression 2-EFA is about 19% at a sample size of 1,500, while the other methods have limited standard error bias ≤ 1%. Panels C and D finds small differences between methods in terms of RMSE and power. For example, RMSE of the moderation effect .04 for the unbiased Bartlett-Regression 2-CFA, .06 for Bartlett-Bartlett 2-CFA, .05 for Regression-Regression 2-CFA at a sample size of 1,500 and the biased Bartlett-Regression 2-EFA has an RMSE of .18. Estimated power of the moderation effect is about 64% at a sample size of 200 for all methods except 19% for the Bartlett-Regression 2-EFA.

In sum, across the investigated factor score implementations, only using the recommended Bartlett scores for Y, and regression scores for X and Z taken from a 2-CFA can recover the main and moderation effects (Devlieger et al. 2016; Skrondal and Laake 2001).

Figure WA17
Follow-up Study 2: Performance Criteria for the Moderation and Main Effects

Panel A: Parameter bias

Moderation effect (β_3) | Main effect of X (β_1) | Main effect of Z (β_2)



Panel B: Standard error bias

Moderation effect (β_3) | Main effect of X (β_1) | Main effect of Z (β_2)

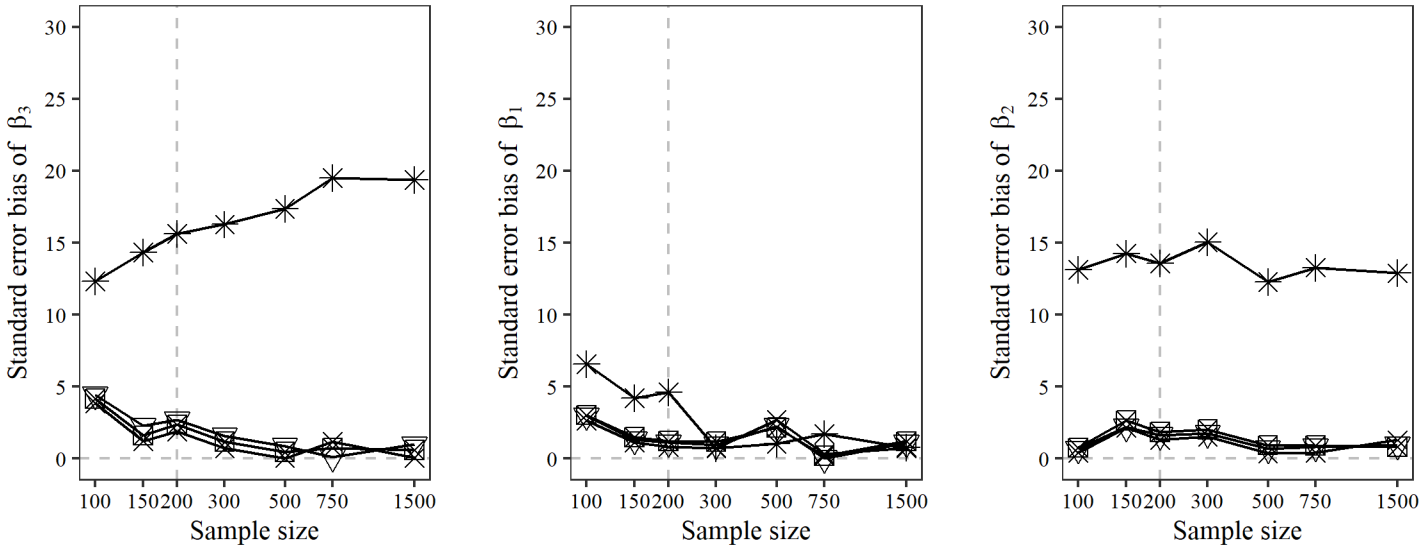
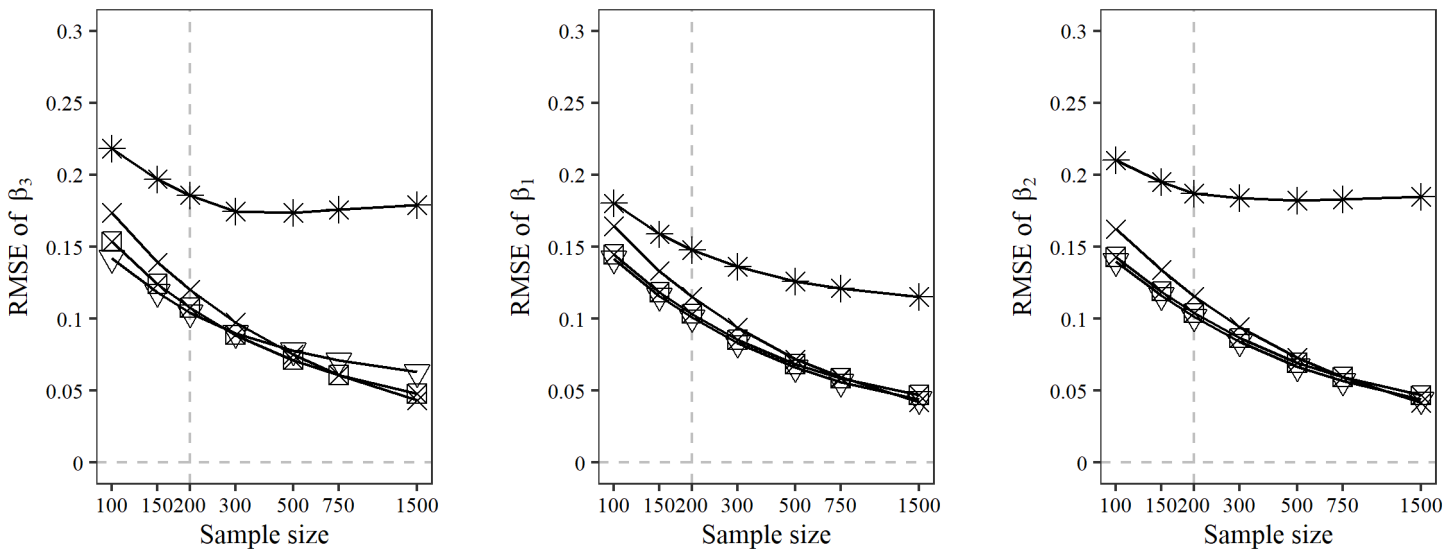
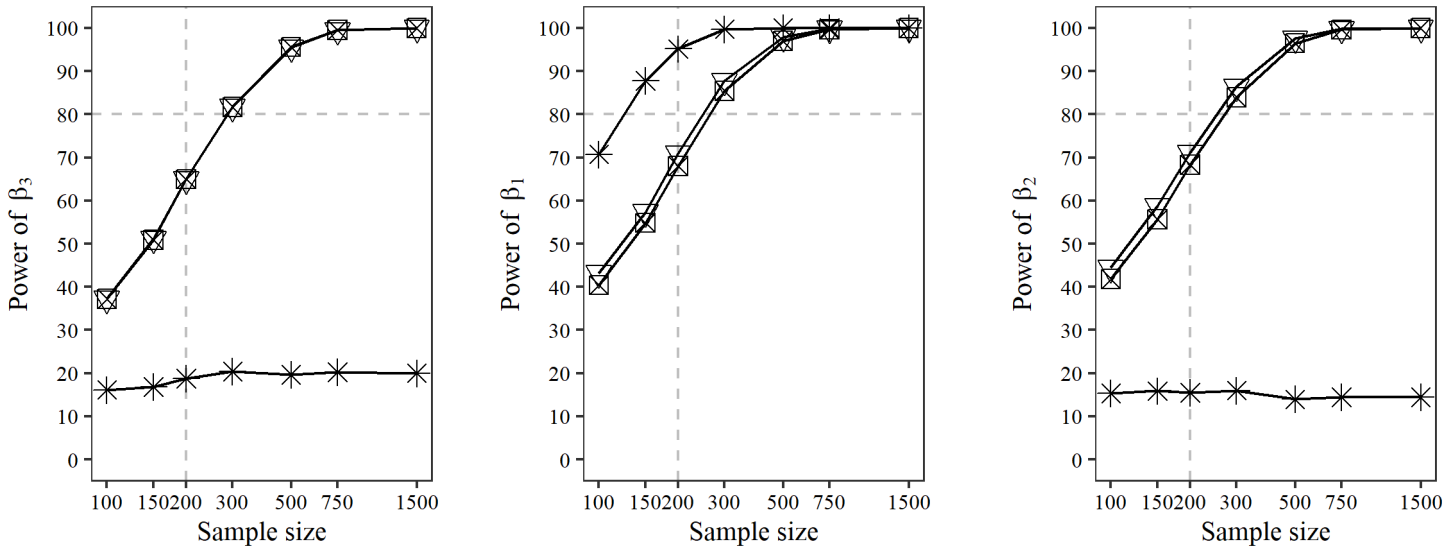


Figure WA17 (CONTINUED)

Panel C: RMSE



Panel D: Power



Legend:

Notes: Plots visualize method parameter bias, standard error bias, root mean squared error (RMSE) and power (as defined in Table 2 of the main text) of the moderation (β_3) and main effects (β_1 and β_2) across sample sizes (log scale). Horizontal dashed lines indicate parameter bias, standard error bias and RMSE of zero and power of 80%. Vertical dashed lines indicate a sample size of 200, about the median in the literature review (Table 1 in the main text).

References of the Web Appendix

Aguinis, Herman, James C. Beaty, Robert J. Boik, and Charles A. Pierce (2005), "Effect Size and Power in Assessing Moderating Effects of Categorical Variables Using Multiple Regression: A 30-Year Review," *Journal of Applied Psychology*, 90 (1), 94-107.

Aroian, Leo A. (1947), "The Probability Function of the Product of Two Normally Distributed Variables," *The Annals of Mathematical Statistics*, 18 (2), 265-71.

Asparouhov, Tihomir and Bengt O. Muthén (2021), "Bayesian Estimation of Single and Multilevel Models with Latent Variable Interactions," *Structural Equation Modeling: A Multidisciplinary Journal*, 28 (2), 314-28.

Auh, Seigyoung, Bulent Menguc, Constantine S. Katsikeas, and Yeon Sung Jung (2019), "When Does Customer Participation Matter? An Empirical Investigation of the Role of Customer Empowerment in the Customer Participation–Performance Link," *Journal of Marketing Research*, 56 (6), 1012-33.

Bishara, Anthony J. and James B. Hittner (2015), "Reducing Bias and Error in the Correlation Coefficient Due to Nonnormality," *Educational and Psychological Measurement*, 75 (5), 785-804.

Bollen, Kenneth A. (1989), Structural Equations with Latent Variables. New York: Wiley.

Cohen, Jacob (1988), Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), "Maximum Likelihood from Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 39 (1), 1-38.

Devlieger, Ines, Axel Mayer, and Yves Rosseel (2016), "Hypothesis Testing Using Factor Score Regression: A Comparison of Four Methods," *Educational and Psychological Measurement*, 76 (5), 741-70.

DiStefano, Christine, Min Zhu, and Diana Mindrila (2009), "Understanding and Using Factor Scores: Considerations for the Applied Researcher," *Practical Assessment, Research & Evaluation*, 14 (20), 1-11.

Eisend, Martin (2015), "Have We Progressed Marketing Knowledge? A Meta-Meta-Analysis of Effect Sizes in Marketing Research," *Journal of Marketing*, 79 (3), 23-40.

Finney, Sara J. and Christine DiStefano (2006), "Non-Normal and Categorical Data in Structural Equation Modeling," in *Structural Equation Modeling: A Second Course*, Gregory R. Hancock and Ralph O. Mueller, eds. Greenwich, Conneticut: IAP.

Fornell, Claes and David F. Larcker (1981), "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," *Journal of Marketing Research*, 18 (1), 39-50.

Fürst, Andreas, Martin Leimbach, and Jana-Kristin Prigge (2017), "Organizational Multichannel Differentiation: An Analysis of Its Impact on Channel Relationships and Company Sales Success," *Journal of Marketing*, 81 (1), 59-82.

Haans, Richard F. J., Constant Pieters, and Zi-Lin He (2016), "Thinking About U: Theorizing and Testing U- and Inverted U-Shaped Relationships in Strategy Research," *Strategic Management Journal*, 37 (7), 1177-95.

Hsiao, Yu-Yu, Oi-Man Kwok, and Mark H. C. Lai (2021), "Modeling Measurement Errors of the Exogenous Composites from Congeneric Measures in Interaction Models," *Structural Equation Modeling: A Multidisciplinary Journal*, 28 (2), 250-60.

Irwin, Julie R. and Gary H. McClelland (2001), "Misleading Heuristics and Moderated Multiple Regression Models," *Journal of Marketing Research* 38 (1), 100-09.

---- (2003), "Negative Consequences of Dichotomizing Continuous Predictor Variables," *Journal of Marketing Research*, 40 (3), 366-71.

Katsikeas, Constantine S., Seigyoung Auh, Stavroula Spyropoulou, and Bulent Menguc (2018), "Unpacking the Relationship between Sales Control and Salesperson Performance: A Regulatory Fit Perspective," *Journal of Marketing*, 82 (3), 45-69.

Kelava, Augustin, Christina S. Werner, Karin Schermelleh-Engel, Helfried Moosbrugger, Dieter Zapf, Yue Ma, Heining Cham, Leona S. Aiken, and Stephen G. West (2011), "Advanced Nonlinear Latent Variable Modeling: Distribution Analytic LMS and QML Estimators of Interaction and Quadratic Effects," *Structural Equation Modeling: A Multidisciplinary Journal*, 18 (3), 465-91.

Kenny, David A. and Charles M. Judd (1984), "Estimating the Nonlinear and Interactive Effects of Latent Variables," *Psychological Bulletin*, 96 (1), 201-10.

Klein, Andreas G. and Bengt O. Muthén (2007), "Quasi-Maximum Likelihood Estimation of Structural Equation Models with Multiple Interaction and Quadratic Effects," *Multivariate Behavioral Research*, 42 (4), 647-73.

Klein, Andreas and Helfried Moosbrugger (2000), "Maximum Likelihood Estimation of Latent Interaction Effects with the LMS Method," *Psychometrika*, 65 (4), 457-74.

Lastovicka, John L. and Kanchana Thamodaran (1991), "Common Factor Score Estimates in Multiple Regression Problems," *Journal of Marketing Research*, 28 (1), 105-12.

Marsh, Herbert W., Zhonglin Wen, and Kit-Tai Hau (2004), "Structural Equation Models of Latent Interactions: Evaluation of Alternative Estimation Strategies and Indicator Construction," *Psychological Methods*, 9 (3), 275-300.

Moosbrugger, Helfried, Karin Schermelleh-Engel, and Andreas Klein (1997), "Methodological Problems of Estimating Latent Interaction Effects," *Methods of Psychological Research Online*, 2 (2), 95-111.

Muthén, Linda K. and Bengt O. Muthén (2019), Mplus User's Guide (8th ed.). Los Angeles, CA: Muthén & Muthén.

Oliveira, Amílcar, Teresa Oliveira, and Antonio Seijas-Macías (2016), "Evaluation of Kurtosis into the Product of Two Normally Distributed Variables," *AIP Conference Proceedings*, 1738 (1), 1-4.

Peterson, Robert A. (1994), "A Meta-Analysis of Cronbach's Coefficient Alpha," *Journal of Consumer Research*, 21 (2), 381-91.

Peterson, Robert A. and William R. Wilson (1992), "Measuring Customer Satisfaction: Fact and Artifact," *Journal of the Academy of Marketing Science*, 20 (1), 61.

Pieters, Rik (2017), "Meaningful Mediation Analysis: Plausible Causal Inference and Informative Communication," *Journal of Consumer Research*, 44 (3), 692-716.

R Core Team (2020), "R: A Language and Environment for Statistical Computing." Vienna, Austria: R Foundation for Statistical Computing.

Raykov, Tenko (1997), "Scale Reliability, Cronbach's Coefficient Alpha, and Violations of Essential Tau-Equivalence with Fixed Congeneric Components," *Multivariate Behavioral Research*, 32 (4), 329-53.

Reinholtz, Nicholas, Daniel M. Bartels, and Jeffrey R. Parker (2015), "On the Mental Accounting of Restricted-Use Funds: How Gift Cards Change What People Purchase," *Journal of Consumer Research*, 42 (4), 596-614.

Rosenthal, R. and M. R. DiMatteo (2001), "Meta-Analysis: Recent Developments in Quantitative Methods for Literature Reviews," *Annual Review of Psychology*, 52 (1), 59-82.

Schermelleh-Engel, Karin, Andreas Klein, and Helfried Moosbrugger (1998), "Estimating Nonlinear Effects Using a Latent Moderated Structural Equations Approach.," in *Interaction and Nonlinear Effects in Structural Equation Modeling*, Randall E. Schumacker and George A. Marcoulides, eds. Mahwah, NJ: Lawrence Erlbaum Associates.

Skrondal, Anders and Petter Laake (2001), "Regression among Factor Scores," *Psychometrika*, 66 (4), 563-75.

Umbach, Nora, Katharina Naumann, Holger Brandt, and Augustin Kelava (2017), "Fitting Nonlinear Structural Equation Models in R with Package nlsem," *Journal of Statistical Software*, 77 (1), 1-20.

Wathne, Kenneth H., Jan B. Heide, Erik A. Mooi, and Alok Kumar (2018), "Relationship Governance Dynamics: The Roles of Partner Selection Efforts and Mutual Investments," *Journal of Marketing Research*, 55 (5), 704-21.

Wooldridge, Jeffrey M. (2015), Introductory Econometrics: A Modern Approach (6th ed.). Boston, MA: Cengage Learning.

Yuan, Ke-Hai and Lifang Deng (2021), "Equivalence of Partial-Least-Squares SEM and the Methods of Factor-Score Regression," *Structural Equation Modeling: A Multidisciplinary Journal*, 28 (4), 557-71.

Yuan, Ke-Hai, Yong Wen, and Jiashan Tang (2020), "Regression Analysis with Latent Variables by Partial Least Squares and Four Other Composite Scores: Consistency, Bias and Correction," *Structural Equation Modeling: A Multidisciplinary Journal*, 27 (3), 333-50.