Abstract

Monte Carlo simulation studies often face generalizability issues due to researchers degrees of freedom in operationalizing general claims, potentially leading to bias, ambiguity, and conflicting results. Most open science practices may not fully address biases inherent in the design stage of simulation studies. This thesis introduces *Adversarial Simulation* adapting adversarial collaboration from empirical research to simulation studies to reduce bias, enhance rigor, and improve generalizability. As a case study, conflicting findings on the performance of Structural After Measurement (SAM) versus traditional Structural Equation Modeling (SEM) were examined. Two collaborators independently replicated prior studies supporting opposing views: one found that SAM outperformed SEM under model misspecifications, especially with small samples and low indicator reliability; the other found that SEM outperformed SAM when negative cross-loadings or residual correlations were omitted. Despite one collaborator withdrawing, a unified simulation study was conducted by the other formally demonstrating the feasability of *Adversarial Simulation*. The joint study revealed that SAM generally offers advantages over SEM, particularly in handling model misspecifications involving both positive and negative omitted cross-loadings and residual correlations. Specifically, SAM demonstrated lower bias and root mean square error in estimating structural parameters, higher convergence rates, and fewer improper solutions compared to SEM under challenging conditions. Constrained SEM methods mitigated some convergence issues but did not fully match SAM's performance. This process demonstrates that adversarial collaboration is a viable approach in simulation studies, effectively reducing ambiguity and possibly bias in design. Despite challenges like increased resource demands and a narrow field of applicability, *Adversarial Simulation* holds promise for enhancing the robustness and transparency of simulation-based statistical research.

Document Version: Generated using AdversarialSimulation in state of Git commit SHA 629526f

Table of contents

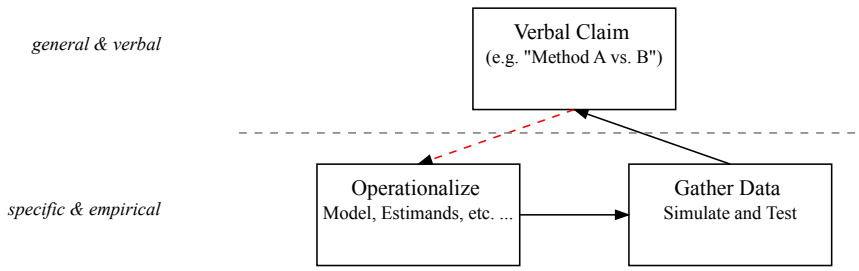
Introduction

The Challenge of Generalizability in Simulation Studies

Karl Popper once described science "as the art of systematic over-simplification — the art of discerning what we may with advantage omit" (Popper, 1988). This pointedly encapsulates the fundamental cycle of empirical research. Researchers formulate general claims about the world, translate them into specific, measurable constructs, select appropriate methods to collect data from specific populations, and finally update their beliefs about these general claims based on the gathered data (Supino, 2012). A core challenge in every research endeavour is this mapping from the general to the specific when designing and conducting a study and, conversely, from the specific and empirical back to the general when interpreting the results. This traversing between the abstract and concrete is the crux of most research, and it is where many, if not all, researchers degrees of freedom lie (Carrillo & Martínez, 2023; Dellsén & Baghramian, 2021; Martínez & Huang, 2011). If these mappings are ambiguous or lack transparency due to explicit or implicit questionable research practices, irreproducible and ungeneralizable results (Camerer et al., 2018; Earp & Trafimow, 2015; McCarley et al., 2023; Nosek et al., 2022) as well as persistent and seemingly unresolvable disagreements on the general and verbal level can emerge (Cleeremans, 2022; Dellsén & Baghramian, 2021). This fundamental challenge of generalizability - i.e. the extent to which research findings apply beyond the specific context of a study (Yarkoni, 2022) - manifests uniquely in Monte Carlo simulation studies. These are widely utilized tools to assess the reliability and validity of statistical methods by testing their performance against controlled, simulated data, where the true population parameter values (of the data generating mechanism) are known. This approach allows researchers to investigate biases, variances, and other properties of estimators or models under various conditions, often informing best practices and model improvements (Banack et al., 2021; Morris et al., 2019; Thomopoulos, 2012). These simulations, constrained by their specific settings, heavily depend on inductive reasoning to draw general conclusions, as it is impossible to simulate and test every conceivable combination of population and analysis model (Feinberg & Rubright, 2016; Gilbert & Miratrix, 2024). Consequently, researchers face numerous degrees of freedom in deciding which prototypical models and settings to examine when assessing a method's general applicability and performance (Bollmann et al., 2015; Kulinskaya et al., 2020). Hence, it becomes clear that simulation studies follow the same general research cycle as empirical studies, facing similar challenges of generalizability and ambiguity, where valid and reproducible mappings between each stage are pivotal for ensuring generalizability (see Figure 1).

Figure 1

The Basic Cycle of Research (in simulation studies)



Note. The dashed red arrow from *Verbal Claim* to *Operationalization* indicates translational validity as the mapping from the general to the specific not directly addressed by most open science practices.

To address these challenges, just like for empirical research, open science practices such as preregistration, open data, and reproducibility have been proposed for simulation studies (O'Kelly et al., 2017; Pawel et al., 2023). These practices are pivotal for increasing rigor and transparency in simulations and beyond. However, in the basic cyclical process of research, as outlined above, they address the issue of generalizability only indirectly by promoting scientific rigor at the point of moving from operationalization to data simulation, e.g. by enforcing transparency and reproducibility of the simulation (Luijken et al., 2023), or at the point of verbally interpreting gathered data through adherence to preregistration protocols. Crucially, however, these practices do not address the transition from the verbal claim to its operationalization (See 1), also referred to as *translational validity* (Clark et al., 2022a; Slife et al., 2016). Decisions about operationalization, for example, which models and simulation settings to choose remain subject to researchers degrees of freedom and their (often implicit) biases (Buchka et al., 2021; Flake & Fried, 2020). This can produce conflicting verbal claims and disagreements that may not be readily resolvable, or only inefficiently so, by independently conducting and publishing simulation studies that may require extensive review articles to inform a broader audience of researchers (Hamaker, 2023). Researchers biases toward specific methods they have developed themselves may further amplify these divergences, affecting not only the interpretation of results but crucially also choices in the design of simulation studies (Buchka et al., 2021).

Adversarial Collaboration

For empirical research, Adversarial Collaboration (AC) has been proposed to address ambiguity and improve generalizability throughout the research cycle in case of entrenched disagreements between researchers and theories. Pioneered by Latham et al. (1988) and Mellers et al. (2001) it has since been recognized for its potential within the open science movement (Clark et al., 2022a; C. Clark & Tetlock, 2021; Rakow, 2022). Unlike other open science practices, which may not account for researchers degrees of freedom in hypothesis generation and operationalization, AC allows for detecting and reducing biased methodological decisions. In AC, opposing researchers first identify general verbal theoretical disputes, agree on a shared research question that could settle the debate, collaboratively design studies they agree to have the potential to change their minds and jointly publish the results regardless of the outcome (Kahneman & Klein, 2009; Latham et al., 1988; Melloni et al., 2021). This process aims to unveil and concretize even subtle discrepancies in methodological assumptions, decisions and in the framing of conclusions (Clark et al., 2022a). By tracing general and verbal disagreements back to their specific empirical origins, it fosters a shared language of assumptions and operationalizations, reducing ambiguity and enhancing generalizability. Hence, it promises to enhance rigor and transparency and, importantly, reduce ambiguity and bias at the operationalization and design stage of studies (C. Clark & Tetlock, 2021).

SAM vs. SEM - A Case Study of *Adversarial Simulation*

In this project, we aimed to adapt this concept of AC from empirical research to Monte Carlo simulation studies, examining the feasibility and viability of such an *Adversarial Simulation*. As a substantive test case, we focused on recent conflicting findings regarding the performance of Structural After Measurement (SAM) — a method for Structural Equation Model (SEM) estimation recently reintroduced by Rosseel & Loh (2022).

Structural equation modelling (SEM) encompasses various statistical techniques frequently applied in the social and behavioural sciences (Bollen, 2014; Hoyle, 2012). SEM is most commonly employed to study models incorporating measurement and structural components. The measurement model describes the relationships between latent variables and their observed indicators, while the structural model specifies the relationships among the latent variables themselves, often reflecting substantive theoretical constructs of interest (Hair Jr et al., 2021). Traditional SEM estimation methods, like maximum likelihood estimation, optimize all parameters of a model simultaneously (under the assumption of multivariate normality) by minimizing a discrepancy function $F(\theta)$, where θ represents all parameters of both the

measurement and structural models (Kline, 2023). While powerful, this system-wide estimation suffers from several shortcomings, such as non-convergence, improper solutions (with solutions including parameters out of their definitional range, such as negative variances (Van Driel, 1978)), poor model fit, and estimation biases arising from local measurement misspecifications that can affect the entire model. They also typically require large sample sizes for adequate performance, especially in complex models with many parameters (Rosseel, 2020). SAM - as proposed by Rosseel & Loh (2022) - addresses these issues and separates the estimation process into two distinct stages. First, the measurement model parameters are estimated independently to capture the relationships between latent variables (η) and their observed indicators (y), represented by:

$$y = \mu + \Lambda\eta + \epsilon, \tag{1}$$

where μ is a vector of intercepts, Λ is the factor loading matrix, and ϵ denotes measurement errors. In the second stage, structural model parameters are estimated using the latent variable estimates from the first stage, modelled as:

$$\eta = \alpha + B\eta + \zeta, \tag{2}$$

with α as a vector of intercepts, B as the matrix of structural coefficients, and ζ as structural disturbances. Rosseel & Loh (2022) proposes two distinct approaches to SAM estimation: (1) Local SAM constructs estimators for latent variable means and covariances from first-stage estimates and applies them directly in second-stage analyses (e.g., linear regression). Expected values $E(\eta)$ and covariance $\text{Var}(\eta)$ are derived as:

$$E(\eta) = M(y - \mu), \tag{3}$$

$$\text{Var}(\eta) = M(S - \Theta)M^T, \tag{4}$$

where M is derived from factor loadings and measurement error covariances, S is the sample covariance matrix of y, and Θ is the covariance matrix of ϵ. (2) Global SAM keeps first-stage measurement model parameters fixed while estimating structural model parameters using standard SEM techniques.

Recent studies by Rosseel & Loh (2022) and Dhaene & Rosseel (2023) have shown that SAM can outperform traditional SEM estimation in the presence of model misspecifications, especially in small to moderate sample sizes. However, Robitzsch (2022) has challenged these findings, arguing that SAM may systematically underestimate parameters in the presence of omitted negative cross-loadings or residual correlations. This disagreement highlights the need for

a systematic evaluation of SAM's performance in the presence of model misspecifications in small to moderate sample sizes. These diverging claims served as the starting point for an AC between a fellow student researcher (Collaborator B, Kosanke) and me (Collaborator A, Kriegmair), each representing one of the above sides of the differing findings. We developed a basic framework to tailor the concept of AC to simulation studies and facilitate a structured and systematic conduct and evaluation of the AC process in which a conceptual replication of respective findings by each collaborator marks the starting point representing a suitable testing ground for *Adversarial Simulation*. The framework, outlined in detail in the following section, consisted of two rounds: In the first round, each collaborator would independently conduct a separate simulation study. The AC would take place in the second round, and we planned to collaboratively design and conduct a study based on the first round. This two-stage approach was designed to systematically highlight differences between the individual approaches and establish a virtual foundation for collaboration before engaging in a joint effort in our case study. As the first step of this process, we jointly formulated two *substantive* research questions based on the above-mentioned conflicting claims:

1. How do SAM and traditional system-wide SEM estimation compare in terms of bias, mean squared error (MSE), and convergence rates in small to moderate samples?
2. What is the impact of model misspecifications, such as residual correlations and cross-loadings, on the performance of SAM compared to traditional SEM methods?

Finally, this case study explored the overarching research question: Is AC a viable and practical approach for resolving conflicting research claims and enhancing generalizability and rigor in simulation studies?

Methods

A Framework for Adversarial Collaboration

We developed a specific adversarial simulation framework and structured the collaboration into two rounds. First, as outlined above, the collaborators identify their disagreement and jointly formulate research questions that could resolve the diverging claims. Based on these research questions, in the first round, each collaborator independently conducts a separate simulation study. In the second round, the collaborators come together to work on a combined study, building on the findings from the first round. This two-step approach is designed to highlight differences in a systematic way and to establish a virtual foundation for collaboration before engaging in a joint effort in our case study. Further, we aimed to adhere the individual studies to a protocol of core steps of a simulation study adapted from Paxton et al. (2001) where critical and distinctive decisions by the researchers occur (see Figure 2). Thus, this was aimed at facilitating a

systematic and stepwise comparision of the individual studies and integration to a joint study.

Table 1

A Framework of Simulation Studies (adapted from Paxton etl al., 2001)

Step	Description
1. Aims & Research Questions	Agreed upon before any Adversarial Collaboration (e.g., examine model fit under misspecification)
2. Population Model Specification	Optional: Define structure (e.g., CFA, SEM), size (number of latent variables and indicators) and complexity (e.g., cross-loadings) of population models.
3. Data Generating Mechanism	Choose between resampling and parametric draw and set the random data generation method.
4. Experimental Design	Define varying factors (e.g., sample size and distribution).
5. Method Selection	Select estimation methods based on the research question.
6. Defining Estimands	Reflect applied values, e.g., R^2, statistical significance, power considerations.
7. Performance Measures	Choose performance metrics (e.g., bias, sensitivity, accuracy); set simulation number for adequate Monte Carlo standard error.
8. Software Selection	Select software, libraries, and functions for simulation.
9. Analysis	Decide on descriptive vs. inferential analysis and performance criteria

Individual Simulation Studies

Studies by Collaborator A (Kriegmair)

The design and implementation of my (Collaborator A) individual simulation studies follow the structure established for the adversarial simulation framework to facilitate stepwise collaboration. They are based on a preregistered protocol (Kriegmair, 2024) but include some deviations from it (See Appendix A for the full protocol and all deviations from the preregistration). As outlined above, in the initial phase of the current case study, I conducted two

simulation studies without my collaborator's involvement aimed at conceptually replicating the findings regarding SAM compared to standard SEM estimation of Rosseel & Loh (2022) and Dhaene & Rosseel (2023) as a basis for further adversarial collaboration with my collaborator. However, there are several differences in the design and setup of the studies compared to the original studies, as outlined below.

Aims, objectives and research questions Both studies aimed to evaluate the performance of traditional SEM with maximum likelihood (ML) compared to global SAM (gSAM), local SAM with maximum likelihood (lSAM-ML), and local SAM with unweighted least squares (lSAM-ULS) under various conditions. The two research questions we jointly established prior to conducting the studies served as a general basis for both studies:

1. How do SAM and traditional system-wide SEM estimation compare in terms of bias, mean squared error (MSE), and convergence rates in small to moderate samples?
2. What is the impact of model misspecifications, such as residual correlations and cross-loadings, on the performance of SAM compared to traditional SEM methods?

Population Models and Data Generating Mechanism

Study 1 Data were generated based on a 5-factor population structural model with three indicators for each factor. Four different models were simulated (see Figure 2). In line with Rosseel & Loh (2022), this model design was chosen to represent a realistic model with sufficient complexity to pose a challenge for the estimation methods, especially in the presence of misspecifications:

- Model 1.1: Correctly specified model.
- Model 1.2: Misspecified with cross-loadings in the population model that are ignored in the estimation model (model 1.1)
- Model 1.3: Misspecified with correlated residuals and a reversed structural path between the third and fourth latent factors in the population model that are ignored in the estimation model (model 1.1)
- Model 1.4: Misspecified with a bidirectional structural relation between factors 3 and 4 specified as only one directional

Factor loadings were fixed across all reliability conditions, with the first indicator of each factor serving as the scaling indicator ($\lambda = 1.0$) and the other two indicators having loadings of 0.7. Indicator reliability levels were manipulated by adjusting the measurement error variances in the Θ matrix. Specifically, the reliability value was set at different levels (low = 0.3, moderate =

0.5 or high = 0.7) to compute the respective error variances on the diagonal of Θ:

$$\Theta^* = \text{Var}(\eta)\Lambda^T \times \frac{1}{r-1} \tag{5}$$

where $\text{Var}(\eta)$ is the factor variance, Λ^T is the transposed factor loading matrix, and r is the reliability level.

To investigate additional possible and realistic scenarios beyond the ones studied by Rosseel & Loh (2022), model 1.3 included a combination of measurement and structural misspecifications as opposed to only measurement misspecifications to introduce an even more severely misspecified model under which SAM methods might perform even better than traditional SEM. Further, model 1.4 included a (not estimated) bidirectional structural relation between factors 3 and 4 as opposed to the unidirectional reversed one. For all models, the population-level values of the structural parameters were set to 0.1.

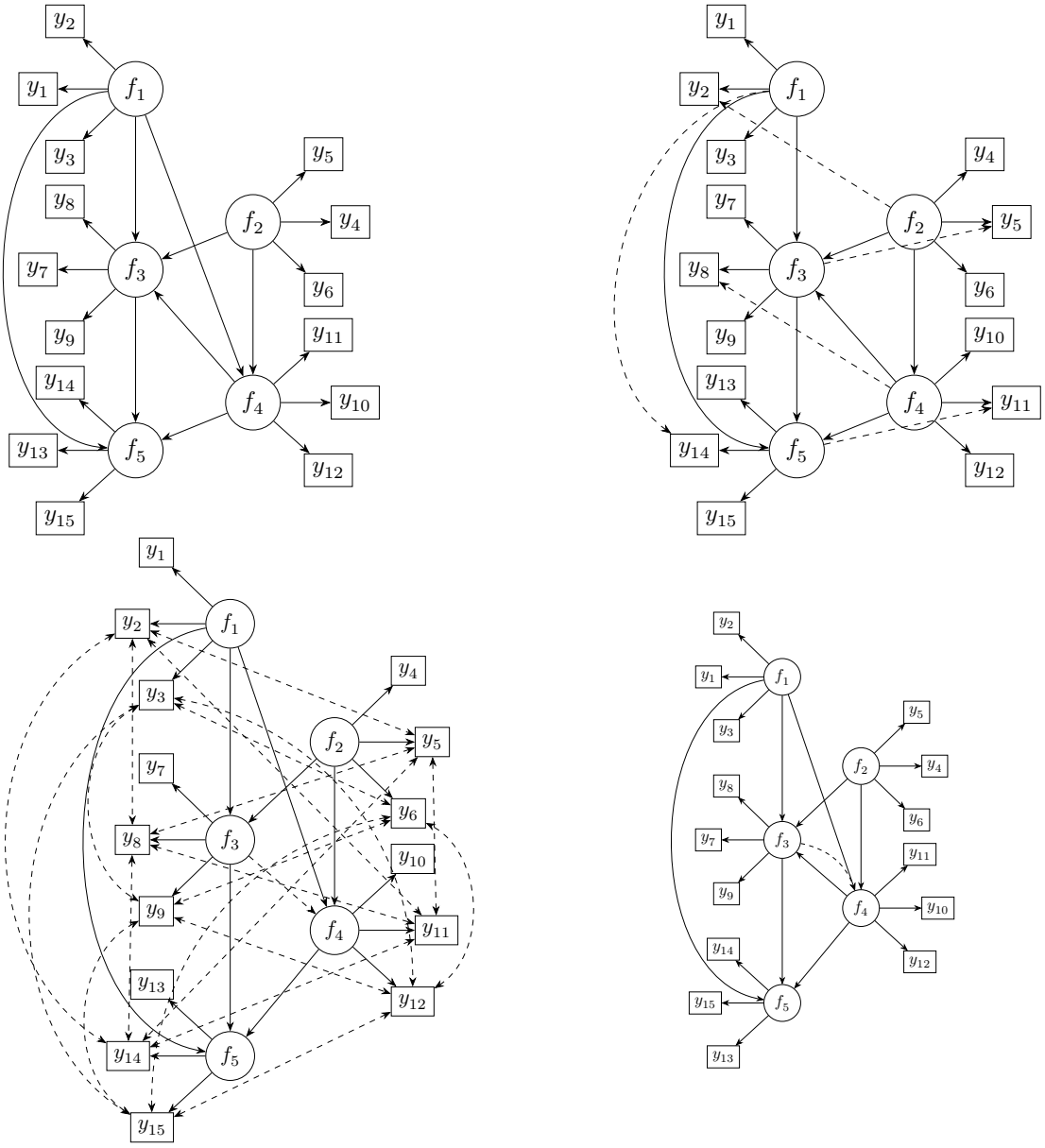
Study 2 Data were generated based on a 5-factor population structural model with three indicators for each factor with loadings set to 1, 0.9 and 0.8 for each factor and reliability modulated like in study 1. Regression weights were set to either 0.183 and 0.224 (low) or 0.365 and 0.447 (medium). This should represent varying variance explained (R^2) by the endogenous factors set at low ($R^2 = 0.1$) or medium ($R^2 = 0.4$). Note, however, that the computation of this was a simplification and does not accurately result in said R^2 values. The aim here was only to generally modulate between lower and higher regression weights. The population models resulted in the following model types with varying misspecification in the estimation model: (1) Structural misspecification with falsely specified paths in the estimation model absent in the population model. (2) correlated residuals and a factor cross-loading in either the exogenous, endogenous part of the model or both with falsely specified paths in the estimation model absent in the population model (see Figure 3). To enable the analysis of the impact of falsely specified paths in the estimation model that are not present in the population model and how well the different methods recover these non-existing relations, both population models included several such misspecifications in addition to the measurement misspecifications evaluated by Dhaene & Rosseel (2023).

Experimental Design

Study 1 Study 1 varied three main conditions: (1) sample sizes of small ($N = 100$), moderate ($N = 400$), and large ($N = 6400$); (2) Indicator reliability of low (= 0.3), moderate (0.5), high (= 0.7); (3) Model specifications: correctly specified model and misspecified with not specified cross-loadings in the population model misspecified with not-specified correlated residuals and a reversed structural path between the the third and the fourth latent factor in the population

Figure 2

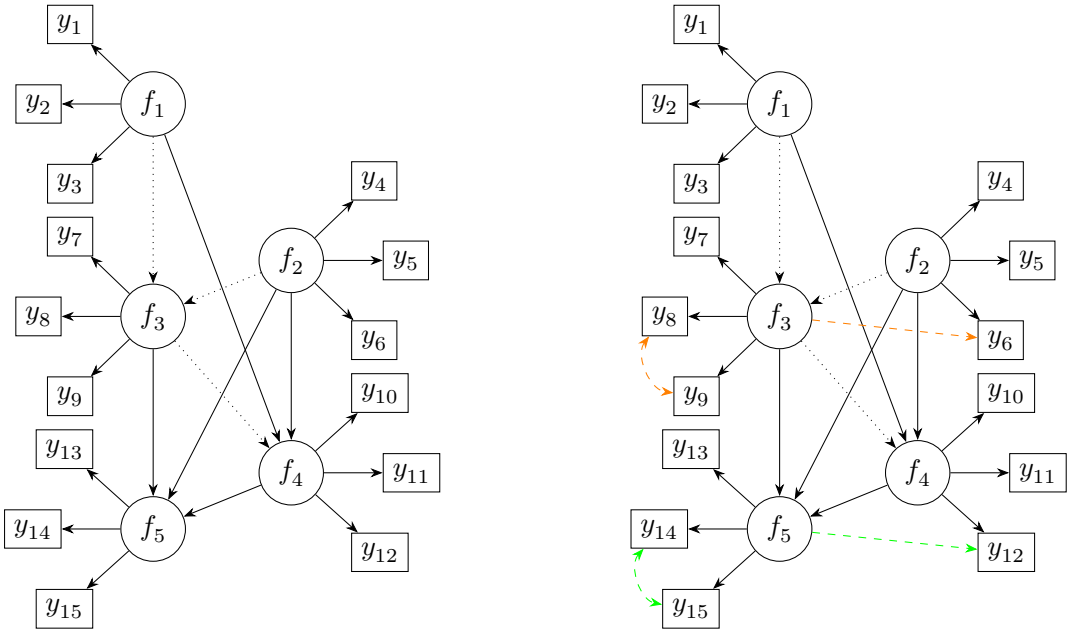
Population Model Variations of Study 1



Note. Error terms are not explicitly shown in the figure. Dashed lines represent relations omitted in the estimation model present in the population model. The figure is adapted from Rosseel & Loh (2022).

Figure 3

Population Model Variations of Study 2



Note. Error terms are not explicitly shown in the figure. Dotted paths represent relations specified in the estimation model not present in the population model. For the model on the right, orange lines represent misspecifications in the exogenous part of the model, and green lines represent misspecifications in the endogenous part. These types of misspecifications result in different realizations of the model when they are modulated as factors of misspecification (endogenous, exogenous or endo- and exogenous) in study 2 but are subsumed under one model here. The figure adapted from Dhaene & Rosseel (2023).

model and a recursive structural relation between factor 3 and 4 in the population specified as only one directional (see Figure 2).

Study 2 Study 2 varied five conditions: (1) sample sizes: small ($N = 100$), medium ($N = 400$), and large ($N = 6400$). (2) Variance explained by endogenous factors: low ($R^2 = 0.1$) and medium ($R^2 = 0.4$). (3) Indicator reliability: low (0.3), moderate (0.5), and high (0.7). (4) Model misspecifications: varying the population model by omitting a residual covariance and a factor cross-loading in different parts of the model. (5) Number of measurement blocks: separate measurement model per latent variable ($b = 5$) and joint measurement model for all exogenous variables ($b = 3$) for the local SAM condition (lSAM-ML).

Method Selection Both studies compared the performance of four estimation methods: Traditional SEM with maximum likelihood (ML), Global SAM with maximum likelihood (gSAM), Local SAM with maximum likelihood (lSAM-ML), Local SAM with unweighted least squares (lSAM-ULS).

Performance Measures For both studies, convergence rates were tracked via lavaan's built-in function, which indicates convergence. Further, improper solutions—converged models that showed negative variances (as the only type of improper solution present)—were tracked via lavaan warning messages. Next, for all converged and proper solutions, bias, for tracking systematic deviations, and root mean squared error for tracking overall accurcy were computed for structural parameter estimates were computed using the following equations (Joshi & Pustejovsky, 2024):

$$\text{Bias} = \bar{T} - \theta \tag{6}$$

$$\text{RMSE} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(T_k - \theta)^2} \tag{7}$$

where T_k is the estimated parameter, $\bar{T}$ is the mean of the estimated parameters, θ is the true parameter value, and K is the number of replications computed. For comparability across varying regression weights for study 2, relative bias and relative RMSE were computed as:

$$\text{Relative Bias} = \frac{\bar{T} - \theta}{\theta} \tag{8}$$

$$\text{Relative RMSE} = \sqrt{\frac{(\bar{T} - \theta)^2 + S_T^2}{\theta^2}} \tag{9}$$

where S_T^2 is the variance of the estimated parameters. Monte Carlo standard errors (MCSE) were computed for bias, RMSE, relative bias, and relative RMSE using the following equations.

$$\text{MCSE}_{\text{Bias}} = \sqrt{\frac{S_T^2}{K}} \tag{10}$$

$$\text{MCSE}_{\text{Relative Bias}} = \sqrt{\frac{S_T^2}{K\theta^2}} \tag{11}$$

$$\text{MCSE}_{\text{RMSE}} = \sqrt{\frac{K-1}{K}\sum_{j=1}^{K}\left(\text{RMSE}_{(j)} - \text{RMSE}\right)^2} \tag{12}$$

$$\text{MCSE}_{\text{Relative RMSE}} = \sqrt{\frac{K-1}{K}\sum_{j=1}^{K}\left(\text{rRMSE}_{(j)} - \text{rRMSE}\right)^2} \tag{13}$$

Software The simulations were executed on a high-performance computing cluster available to the Max Planck Institute for Human Development Berlin (MPIB). All analyses were conducted in R (version 4.4). (R Core Team, 2023). Main libraries included `lavaan` (Rosseel, 2012) for estimation and data generation, `furrr` (Vaughan & Dancho, 2022) for parallelization and `tidyverse`(Wickham et al., 2019), `ggplot2` as well as `kableExtra` (Zhu, 2024) for results analysis and display. To ensure reproducibility and prevent seed synchronization in parallelized simulation runs, a pre-generated list of seeds was applied across all replications, and the simulations were containerized with Docker (Merkel, 2014). Further details, full replication instructions, and a complete list of libraries and dependencies are available in Appendix D or on GitHub.

Analysis and Interpretation Similar to the studies by Rosseel & Loh (2022) and Dhaene & Rosseel (2023), results were interpreted by descriptively comparing the performance measures of the different estimation methods under varying sample sizes, indicator reliability levels, and model misspecifications. Performance metric values were aggregated across all parameters, excluding the misspecified parameters (present in the population but not in the estimation model).

Studies by Collaborator B (Kosanke)

Here Collaborator B's (Kosanke) studies are presented verbatim from his report (Git commit SHA 3e7f706): *The structure of this section closely aligns to our agreed upon structure of simulation studies […]. In a first step, I published a simulation protocol containing all the planned analysis to be replicated from the original paper by Robitzsch (2022). This protocol can be accessed here:*
https://github.com/lkosanke/AdversarialSimulation/blob/main/LK/simulation_protocol.pdf.

Aims, objectives and research questions *For my individual study, I replicated parts of Robitzsch (2022) that were relevant to our two substantive research questions. Overall, I conducted 6 simulation studies.*

Population Models and Data Generation Mechanisms *The most important details with regards to the population models and data-generating mechanisms are visible in Table 7. With regards to the population models, all factors in all studies loaded onto 3 indicators each. I chose*

the population values to align with the original paper by Robitzsch (2022). The multivariate normally distributed data was generated parametrically, based on a specified population model. All simulations were conducted using seeds to allow for the reproducibility of results. For more details on the exact values of each study, see the simulation scripts in the Github repository.

Figure 4

Overview of Simulation Studies Conducted by Kosanke

Study	Model	Correct model included?	Unmodelled RC	Unmodelled CL	N Sizes	φ/ β	λ
Study 1	2-factor-CFA	Yes	1 and 2, both pos. and neg.	x	7	φ = 0.6	Fixed
Study 1b	2-factor-CFA	Yes	x	x	2	φ = 0,2 - 0.8	Varied
Study 2	2-factor-CFA	x	x	1 and 2, both pos. and neg.	7	φ = 0.6	Fixed
Study 3	2-factor-CFA	x	1, pos.	1, pos.	7	φ = 0.6	Fixed
Study 4	5-factors	Yes	20, all pos.	5, all pos.	7	β = 0.1	Fixed
Study 4a	5-factors	x	20, all pos.	5, all pos.	7	β = 0.1 - 0.4	Fixed

Note. Φ: factor correlation, N: sample size, λ: factor loading, σ: residual variance, τ: factor variance, RC: residual correlations, CL: cross-loadings, CFA: confirmatory factor analysis, β: regression coefficient between factors.

Experimental Design of simulation procedures *Overall, 3 different types of factors were varied that can be deduced from Table 7 and are detailed again in the simulation scripts provided. Firstly, I varied the sample size in all studies, ranging from N = 50 to 100.000. I included a smaller sample size N=50 for all studies, to be able to answer our substantive research questions in more detail. Study 1b explicitly investigated the small sample bias of LSAM estimation in low sample sizes. Thus, only N=50 and N=100 were present in this study. Additionally, I varied the amount of misspecification in all studies, either via different numbers of unmodelled residual correlations, cross-loadings, or both. Thirdly, in Studies 1b and 4a, I varied the population values for three model parameters (phi, beta and/ or lambda). Besides studies 1 and 2, I implemented full factorial designs. In Studies 1 and 2 I omitted conditions were both one positive and one negative value would be present. I hypothesize that this was done in Robitzsch (2022) to avoid cancellation of biases, but the authors did not give reasoning for this decision themselves. In Studies 4 and 4a I investigated the differential performance of the estimators in a model that*

included a non-saturated structural model (i.e. regressions between some of the factors). These studies were replications not only of the paper by Robitzsch (2022), but of the first paper on the SAM approach by Rosseel & Loh (2022). In contrast to the other studies, studies 4 and 4a differed in the way the misspecification variation was labelled in Robitzsch (2022). Instead of varying a factor misspecification as in the previous study, they varied 3 different data-generating mechanisms (DGM's) as a whole. Thus the conditions are labelled differently: DGM 1 contained no misspecification. DGM 2 contained 5 cross-loadings in the data-generating model, that were not modelled in the estimated models. DGM 3 contained 20 residual correlations that were not modelled in the models. I extended them to investigate the interaction of beta and N for the 5-factor regression model, as this again was of interest four our substantial research questions. Additionally, I omitted the inclusion of DGM 1 in Study 4a, as it neither contained misspecification (which is central to our research question), nor did it lead to interesting results in the original study.

Method Selection *In terms of estimation methods, I used constrained SEM maximum-likelihood (SEM-ML) and unweighted-least-squares estimation (SEM-ULS), so that loadings and variance parameters were given the constraints that they had to be positive and larger than 0.01. Additionally, I implemented local-SAM (LSAM) and global-SAM (GSAM) estimation, in both maximum-likelihood (LSAM-ML/ GSAM-ML), and unweighted-least-squares estimation (GSAM-ML/ GSAM-ULS) contexts. Exceptions were studies 1b, 4 and 4a, where only LSAM was investigated, as results did not really differ between the two different SAM-methods (Robitzsch, 2022).*

Performance Measures *I calculated the bias and RMSE of the estimated factor correlations in all studies, as well as the standard deviation of the one factor correlation present in Studies 1,2 and 3. For the type of bias calculated, I oriented on Robitzsch (2022), besides in Study 1b. Thus, I calculated average relative bias in Studies 1, 2 and 3, and average absolute bias in Studies 1b, 4 and 4a. In Study 1b, I took the absolute value to see if negative and positive biases canceled each other out in the original study for conditions with lower phi values. In addition to what was done in Robitzsch (2022), I calculated confidence intervals for the bias estimates, but omitted them in the results tables for presentation purposes. The exact computation of the performance measures is detailed in the simulation scripts and results.pdf file in my sub-folder of the Github repository. I did not include a detailed mechanism to capture model convergence as detailed in the first substantive research question. As Robitzsch (2022) argued in their paper, and was shown already in other simulations, using constrained maximum likelihood estimation should resolve convergence issues of classical maximum likelihood estimation in smaller samples (Lüdtke et al., 2021; Ulitzsch et al., 2023). I did include, however, a mechanism to track the total number of warnings for each*

estimation and compare it to the total number of estimations as a sanity check.

Software *All analyses were conducted in R (R Core Team, 2023). I used the packages lavaan, purrr, tidyverse, furrr to conduct the simulations, as well as knitr and kableExtra for presenting the results (Rosseel, 2012; Vaughan & Dancho, 2022; Wickham et al., 2019; Wickham & Henry, 2023; Xie, 2024; Zhu et al., 2024).*

Analysis and Interpretation plan *For the interpretation of results, I oriented on cut-offs that were used in the original paper by Robitzsch (2022). For bias, I interpreted differences of 0.05 or higher as substantial. For SD, I explicitly mentioned percentage reductions of more or equal to 5%. For RMSE, the same interpretation was used for differences of 0.03 or higher. The simulation was repeated 1500 times for each Study.*

Combined Simulation Study

The combined simulation study was conducted based on the individual studies of Collaborator A and Collaborator B to reconcile remaining conflicting claims by unifying divergent study designs and settings. As outlined in the next section reporting the results of the AC process this unified study was conducted without the direct involvement of Collaborator B (Kosanke) as he chose to terminate the case study on AC after the initial stage of individual (replication) studies.

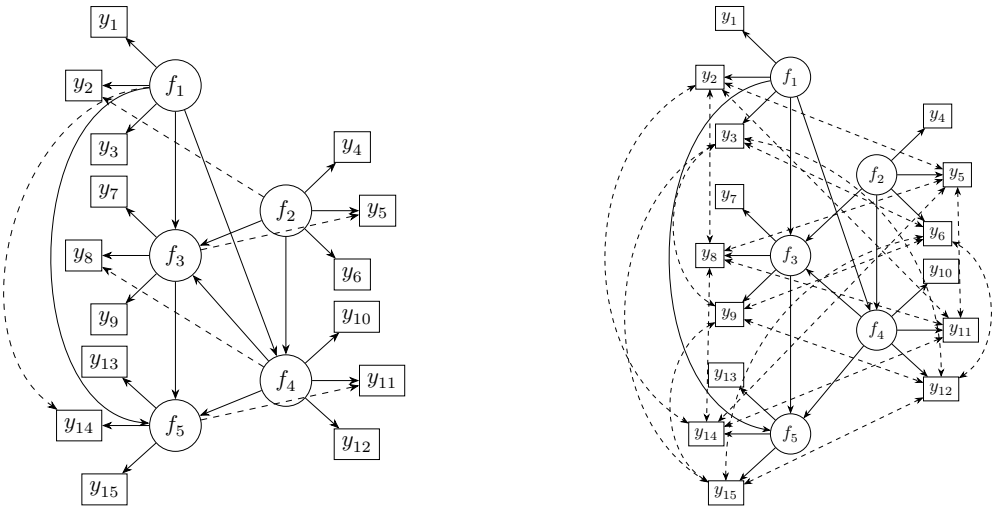
Aims, objectives and research questions Following our framework for collaboration the research questions for the joint study remains the same as specified prior to the individual studies.

Population Models and Data Generation Mechanisms As in my individually conducted studies 1 and 2 (Collaborator A, Kriegmair) data for this *joint* study was generated based on a 5-factor population structural model with 3 indicators for each factor. Factor loadings and indicator reliability was computed in the same way as in my first study. Two different population models were simulated, which resulted in misspecifications of either omitted crossloadings (model 3.1) or omitted correlated residuals (model 3.2). The population-level values of the structural parameters were set to 0.1. Reliability levels were manipulated as in my first study by adjusting the measurement error variances (instead of Kosanke's approach of factor loadings modulation) to achieve a more valid representation of item reliability as the amount of indicator variance explained by the latent factor. The omitted crossloadings (see Figure 6) could either be all positive or negative and were set to be 10% lower in absolute values than the factor loadings. Correlated residuals were also either all positive or all negative and were set to not exceed a factor of 0.6 of the residual variances of the indicators. Thus, this represents a sufficiently complex

model with directed structural paths of interest as a prototypical scenario for which SAM promises to be advantageous, including negative misspecified measurement parameters to test the robustness of SAM to such misspecifications. No CFA models, as in Kosanke's (Collaborator B) studies, were included as SAM is intended to be applied to models with a directed structural part of interest in the presence of misspecifications (Dhaene & Rosseel, 2023; Rosseel & Loh, 2022).

Figure 5

Population Model Variations for Study 3



Note. Error terms are not explicitly shown in the figure. Dashed lines represent relations omitted in the estimation model present in the population model. Unspecified cross-loadings and correlated residuals could be either positive or negative, resulting in two modulations of models 3.1 and 3.2 in the study. The figure is based on Rosseel & Loh (2022).

Experimental Design of simulation procedures The joint study varied three conditions: (1) sample sizes of very small ($N = 50$), small ($N = 100$), moderate ($N = 400$) and large ($N = 6400$). (2) Indicator reliability of low (= 0.3), moderate (0.5) or high (= 0.7); (3) Model misspecifications with not-specified cross-loadings in the population model that were positive or negative (see figure) or not-specified correlated residuals in the population model that were positive or negative (see Figure 4). Thus, negative misspecifications were here included in a more complex model with directed structural paths of interest, modulating reliability as before but with a more comprehensive (lower) range of sample sizes as in my studies.

Method Selection Four estimation methods were compared in this study: bound SEM-ML (with factor and residual variances constrained to be positive), unbound SEM-ML, gSAM (also with ML estimation of the structural model) and lSAM-ML. The choice of these methods was based on the results of the individual studies to (1) observe the effect of constraining the analysis model for standard SEM and (2) directly compare this to unbound standard SEM estimation and

the SAM methods. To limit the computational scope and narrow down the comparison, SAM-ULS and SEM-ULS were not included as ULS estimation is mainly aimed to provide robust estimates in conditions of non-normal data distribution (Krijnen, 1996), which were not simulated in this study. In small samples ULS might still be superior to ML but, based on our previous results, not expected to outperform SAM-ML in this study likely due to the high factor loadings simulated in the populations(Wolins, 1995).

Performance Measures The bias and RMSE of the estimated factor correlations were calculated as in the individual studies and averaged (using absolute values) over all parameters in one model for each condition. Further, to better investigate a potential negative bias that Kosanke (based on Robitzsch (2022)) was assuming for SAM in the presence of negative measurement misspecifications, bias values were analyzed parameterwise to investigate negative bias values without cancellation due to averaging.

Software As in the individual studies by both collaborators, the simulation was conducted in R (version 4.4) (R Core Team, 2023). The same (parallelizable and dockerized) setup as in my studies was used with a pre-generated set of seeds for reproducibility. The simulation scripts are available on GitHub.

Analysis and Interpretation The analysis was conducted largely in the same way as in the individual studies with the adddition of a display of parameter-wise bias values and a direct difference between metrics of SEM and SAM.

Results

Individual Simulation Studies

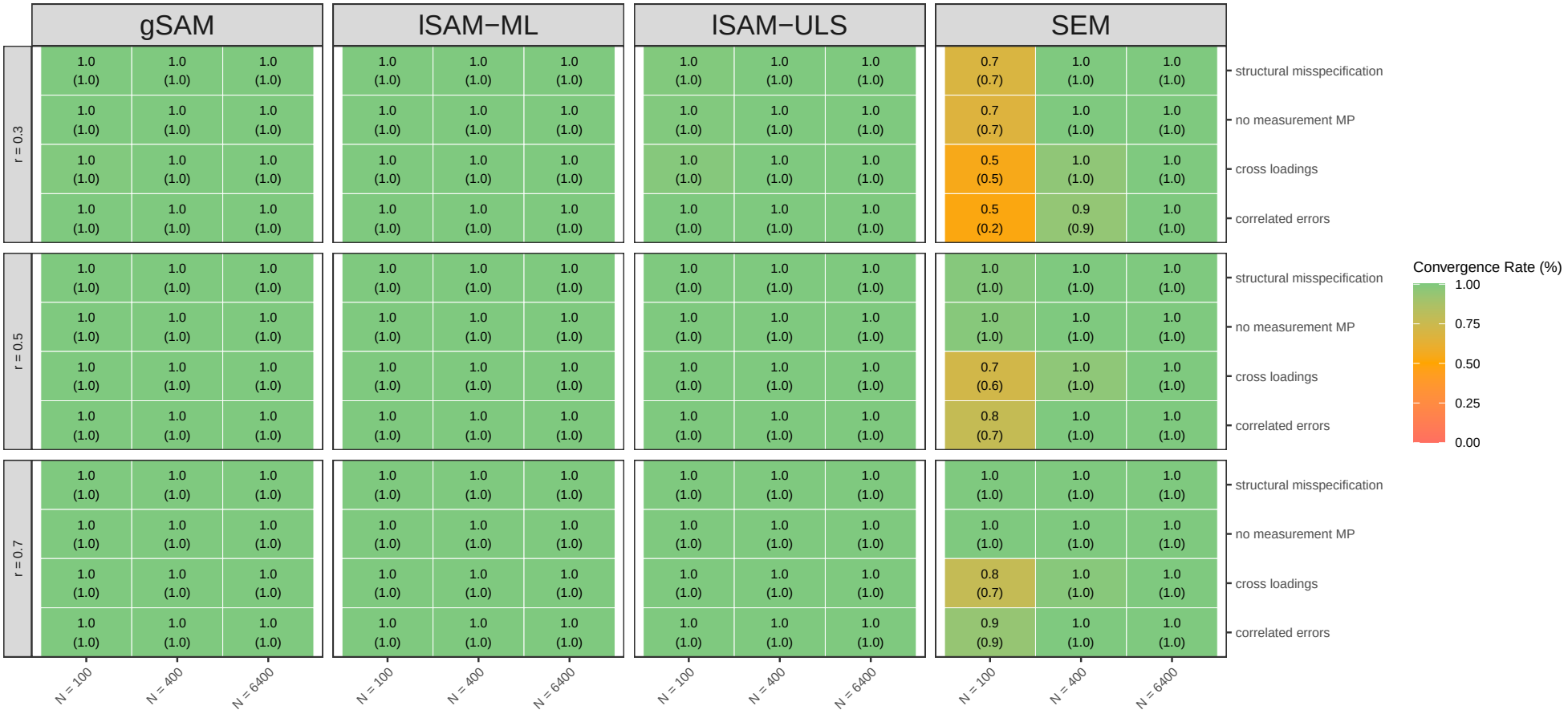
Results of Collaborator A (Kriegmair)

There were no convergence issues for all SAM methods (gSAM, lSAM ML, and ULS) with a convergence rate of 100% and no improper solutions across all conditions, even in small samples with low reliability. Standard SEM demonstrated severe convergence issues, particularly in small samples with low to moderate reliability. The convergence rate was as low as 50%, with 50% of the solutions being improper, especially under the challenging condition of cross-loading misspecification (see Figure 6). Note that all figures referred to in the result sections by Collaborator A (Kriegmair) are based on the report by Dhaene & Rosseel (2023) and modified and adpated for this thesis. Next, the bias of the path coefficient estimates averaged across each model in absolute values showed that in small to medium sample sizes with low to moderate reliability, SAM methods were mostly closer to the true parameters than standard SEM. This

difference was especially pronounced under omitted cross-loadings in the analysis model. However, under correlated residuals, standard SEM was slightly less biased. Large sample sizes and high reliability conditions showed the least bias overall, with no differences observed between the methods (see Figure 7). Further, among the different SAM methods, there was no difference between gSAM and lSAM-ML, while lSAM-ULS performed slightly worse. This pattern was mostly consistent with the RMSE of the path coefficients: SAM methods showed lower values than standard SEM in small to medium sample sizes with low to moderate reliability, indicating higher overall accuracy for SAM methods in challenging conditions. This was notably also the case under correlated residuals where SEM was less biased, which highlights SAM's advantage here as well in light of a trade-off between (slightly higher) bias and precision (see Figure B2). In contrast to the bias, the RMSE showed that lSAM-ML performed better than gSAM and lSAM-ULS under cross-loading and structural misspecifications. However, SAM methods, even though outperforming standard SEM under omitted cross-loadings, still showed substantial deviations in this condition (with bias values between 69% and 77% of the true value for SAM-ML) and inaccuracy (with RMSE values between 86% and 277% of the true value for SAM-ML). Additionally, while increased sample size led to lower RMSE values for all methods, bias only decreased for standard SEM in larger samples. In contrast, SAM methods showed a slight increase in bias with larger samples with measurement misspecifications in low and moderate reliability.

Figure 6

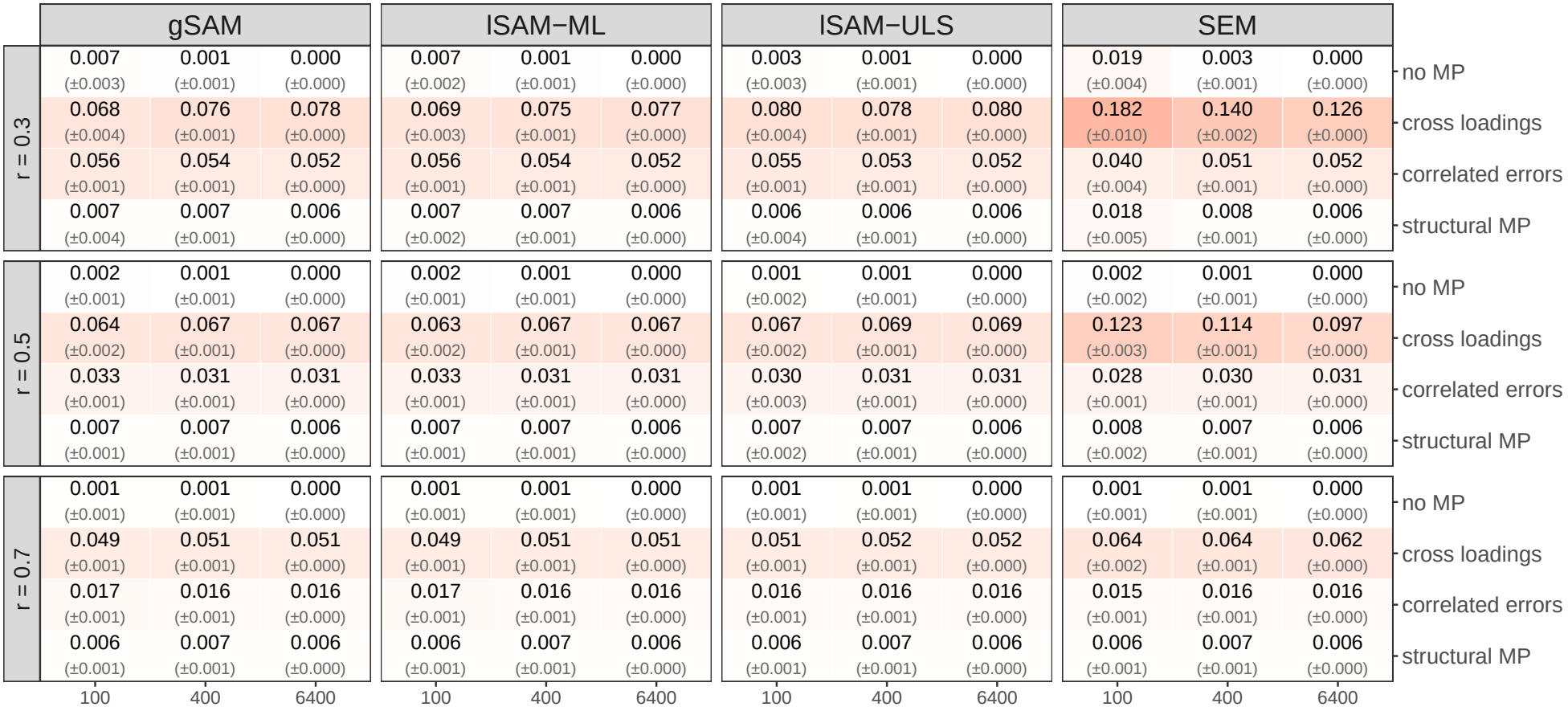
Convergence Rate and Rate of Proper Solutions in Study 1



Note. Convergence and proper solutions (in parentheses) rates across sample sizes (N), reliability (r), and model misspecifications for global SAM (gSAM), local SAM with Maximum Likelihood (lSAM-ML), Unweighted Least Squares (lSAM-ULS), and SEM. Graphic adapted from Dhaene & Rosseel (2023).

Figure 7

Mean Average Bias of Regression Parameters in Study 1



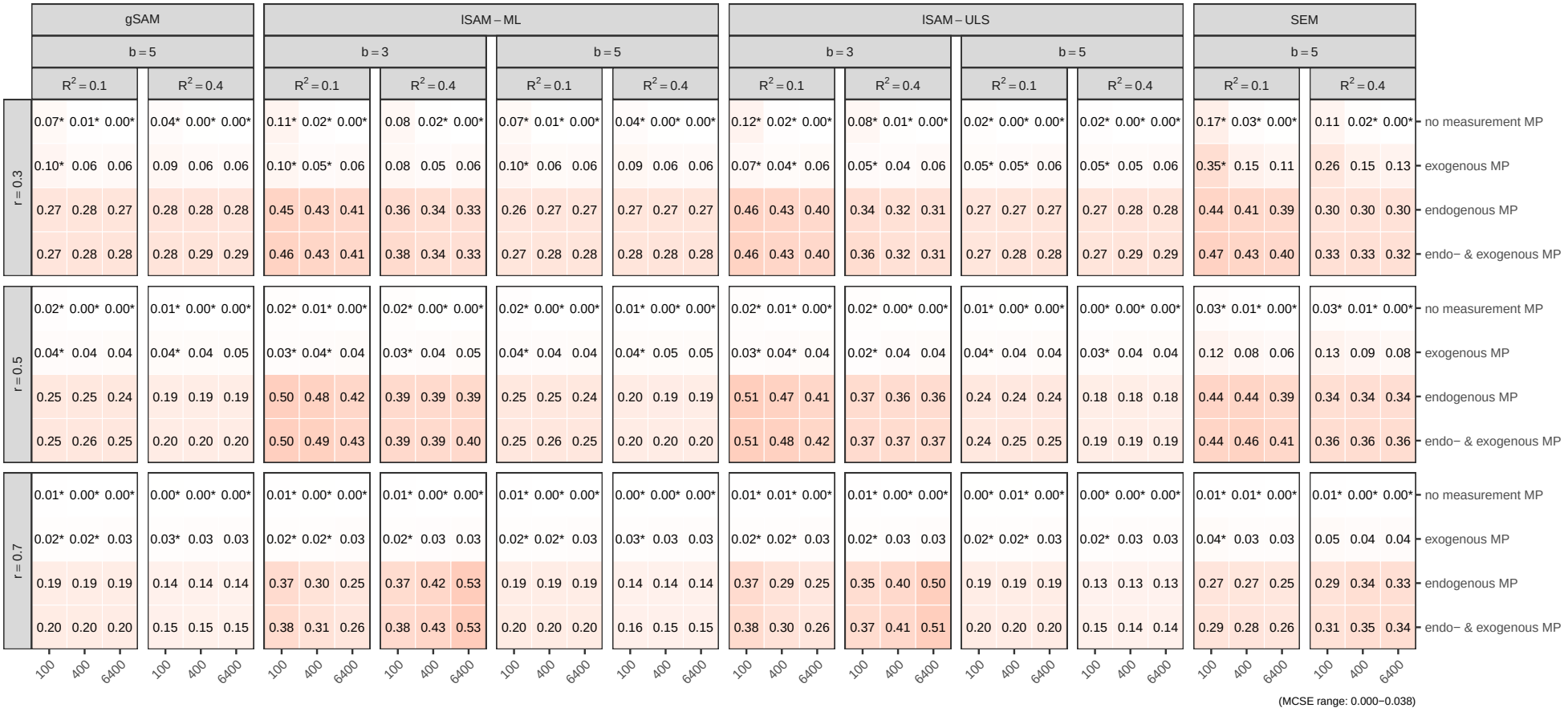
Note. Mean absolute bias averaged (in absolute values) over all parameters with a true value of 0.1 in one model for sample sizes (N), reliability (r), and misspecifications for global SAM (gSAM), local SAM with Maximum Likelihood (lSAM-ML), Unweighted Least Squares (lSAM-ULS) and SEM. Monte Carlo Standard Errors (MCSE) are shown in parentheses for each value. Graphic adapted from Dhaene & Rosseel (2023).

The pattern of results from study 1 was consistent with the findings of study 2, with some additional insights regarding the study-specific conditions. Firstly, as in study 1, there was a 100% convergence rate and rate of proper solutions for all SAM methods across all conditions, even in small samples with low reliability. Standard SEM, in contrast, showed severe convergence issues and frequent improper solutions in small samples with low reliability, with exogenous misspecifications being more challenging than endogenous misspecifications (see Figure B3).

Further, in study 2, the relative bias (to account for modulated path coefficients) of the correctly specified path coefficient estimates averaged across each model in absolute values showed again that in small to medium sample sizes with low to moderate reliability, all SAM estimations were closer to the true parameters than standard SEM. This increased performance of SAM was present only for gSAM and lSAM across all item reliability levels with separate measurement blocks for each factor (b = 5), indicating that joining measurement models in lSAM for exo- and endogenous factors (b = 3) was disadvantageous. All methods performed worse for lower variance explained by the structural model in low and moderate reliability and measurement missepcifcations except SAM methods (with b = 5 and gSAM). The average relative RMSE values of the path coefficients paint a similar picture. Here, too, lower R^2 values were more challenging. Other than for the bias, exogenous misspecifications were more challenging than endogenous misspecifications. Further, gSAM and lSAM-ML with five measurement blocks (here, not ULS) produced notably lower RMSE values than standard SEM; however, only in small and medium samples with low item reliability present. As all population models included structural misspecifications in study 2, there was also a slight advantage visible in low sample size and low reliability without measurement misspecification, indicating that SAM methods are also more robust to the impact of falsely specified paths not present in the population on the estimation of the remaining correctly specified parameters. Figure B6 shows the absolute bias of the parameters misspecified (excluded from the results presented above to avoid distortion of relative metric values by parameters with a true value of 0), indicating that also such parameters are recovered more accurately by SAM methods than standard SEM but only if measurement misspecifications are present.

Figure 8

Mean Average Relative Bias of Regression Parameters in Study 2



Note. Mean relative bias averaged (in absolute values) over all parameters in one model for sample sizes (N), reliability (r), and misspecifications for global SAM (gSAM), local SAM with Maximum Likelihood (lSAM-ML), Unweighted Least Squares (lSAM-ULS) and SEM. * indicating Monte Carlo Standard Error (MCSE) above 10% of the estimate. Graphic adapted from Dhaene & Rosseel (2023).

Overall, the results indicate that SAM methods (gSAM, lSAM-ML, and lSAM-ULS) outperformed standard SEM under challenging conditions. SAM methods achieved a 100% convergence rate with no improper solutions, even in small samples with low reliability, whereas standard SEM exhibited severe convergence issues and high rates of improper solutions, particularly with cross-loading misspecification. SAM methods provided less biased and more accurate path coefficient estimates, especially in small to medium samples with low to moderate reliability and measurement misspecifications like omitted cross-loadings. Among the SAM variants, gSAM and lSAM-ML generally performed better than lSAM-ULS, with lSAM-ML showing superior accuracy under cross-loading and structural misspecifications. Further, separate compared to joint measurement models in lSAM for latent variables was advantageous. Additionally, SAM methods in Study 2 were more robust to structural misspecifications, providing more accurate estimates of correctly specified parameters even when incorrect paths were included in the model.

Results of Collaborator B (Kosanke)

The results of Kosanke's individual simulation studies are presented verbatim from his report (Git commit SHA 4d0e95e):

The full result analysis for my individual study is available here: https://github.com/lkosanke/AdversarialSimulation/blob/main/LK/results.pdf. The repository readme.md contains a detailed explanation of how the analyses were implemented and how they can be reproduced. In this section, I will focus on the most important results only. For the most part, results from Robitzsch (2022) have been successfully replicated: I did not observe substantial convergence issues in any study. Across studies, as in the original paper, SAM did not generally outperform SEM in small to moderate samples. SAM exhibited a negative small sample bias that made SAM appear superior in conditions with unmodelled positive cross-loadings and residual correlations. This bias was especially strong for lower lambda and higher phi or beta values. Going ahead of what was investigated in Robitzsch (2022), I found that this bias is also present in models with lower phi or beta values. Thus, it cannot be concluded that SAM is more robust in models with non-saturated structural parameters. If there was no misspecification or unmodelled negative cross-loadings and residual correlations, SAM tended to perform worse than traditional SEM, as far as can be concluded from my results.

Convergence *As Robitzsch (2022) argued in their paper, I did not expect convergence issues due to constrained ML estimation that only allows for positive variances and loadings. Nevertheless, I captured all messages, warnings and errors that occurred during the simulations. No messages and errors were present in any of the studies. Multiple warnings were observed in the first 4*

simulations, some of them referring to potential problems with convergence. Overall, the number of these warnings was very small compared to the total number of estimations performed. They amounted to between 0.5-1.8%. In studies 4 and 4a, an even smaller number of warnings was present, amounting to problems in 0.02% of estimations in study 4 and 0.1% in study 4a. These warnings referred to potential problems with positive definite matrices and model identification. In total, these numbers are negligible in size and align with the report of Robitzsch (2022), that convergence issues were not substantial for my estimations. Additionally, a larger number of warnings was present with regards to the computation of fit indices in these final two studies. As we were not interest in fit indices in our research question, they were not relevant for our purposes. A detailed analysis of all the warnings was conducted in the results.pdf document in my sub-folder of the Github repository.

Conditions without misspecification *Tables 2 and 3 show the most relevant results of Studies 1 and 4 where I investigated the comparative performance of SAM vs. traditional SEM estimation under correctly specified models. Here it became apparent, that in absence of misspecification, none of the two estimation methods clearly outperformed the other. In Study 4a, only slight differences could be observed in terms of bias and RMSE between LSAM- and classical ML-estimation. In Study 1, both SEM outperformed all SAM estimators in samples of N=50-500. This was true for both relative bias and RMSE, and visible for the former in Table 3. Here, SAM's negative small sample bias is already visible as well.*

Table 2

Study 4 (Kosanke): RMSE for DGM 1 (without misspecification).

	Sample Size						
Method/Metric	**50**	**100**	**250**	**500**	**1000**	**2500**	**100000**
SEM ML	0.188	0.123	0.075	0.051	0.037	0.023	0.004
SEM ULS	1.062	0.128	0.077	0.053	0.037	0.023	0.004
LSAM ML	0.165	0.115	0.072	0.050	0.036	0.023	0.004

Note. SEM ML = Maximum-likelihood estimation, SEM ULS = Unweighted-least-squares estimation, LSAM ML = Local-SAM-maximum-likelihood estimation.

Table 3

Study 1 (Kosanke): Relative bias in conditions without misspecification.

	Sample Size						
Method/Metric	**50**	**100**	**250**	**500**	**1000**	**2500**	**100000**
SEM ML rel bias	-0.045	-0.011	0.001	-0.003	0.003	-0.002	-0.000
SEM ULS rel bias	0.024	0.022	0.012	0.002	0.006	-0.001	-0.000
LSAM ML rel bias	-0.394	-0.270	-0.111	-0.056	-0.022	-0.011	-0.000
LSAM ULS rel bias	-0.393	-0.270	-0.111	-0.056	-0.022	-0.011	-0.000
GSAM ML rel bias	-0.394	-0.270	-0.111	-0.056	-0.022	-0.011	-0.000
GSAM ULS rel bias	-0.393	-0.270	-0.111	-0.056	-0.022	-0.011	-0.000

Note. SEM ML = Maximum-likelihood estimation, SEM ULS = Unweighted-least-squares estimation, LSAM ML = Local-SAM-maximum-likelihood estimation, LSAM ULS = Local-SAM-unweighted-least-squares estimation, GSAM ML = Global-SAM-maximum-likelihood estimation, GSAM ULS = Global-SAM-unweighted-least-squares estimation.

Conditions with negatively valenced unmodelled parameters *Studies 1 and 2 explicitly investigated negatively valenced unmodelled parameters in the generating model. In these studies, it became apparent that traditional SEM outperformed SAM estimation. As can be seen in Table 4, both SEM estimators outperformed all four SAM estimators in terms of relative bias with two negative residual correlations present. The same was true in Study 2, in the presence of two negative cross-loadings. In both these cases, bias values overall remained high but substantially less so in the traditional SEM methods. when comparing them in small to moderate sample sizes. Additionally, slight differences between the two approaches arose in these two examples in terms of RMSE, as can be seen in Table 5 for the negative cross-loadings in study 2.*

Conditions with positively valenced unmodelled parameters *In terms of performance for positively valenced cross-loadings and residual correlations, SAM appeared to outperform traditional SEM estimation, but not in all scenarios of interest. Table 6 shows this finding in Study 3, in conditions with both one unmodelled residual correlation and one cross-loading. Only from N=100-1000 did SAM outperform SEM.*

Table 4

Study 1 (Kosanke): Relative bias in conditions with two negative unmodelled residual correlations.

	Sample Size						
Method/Metric	**50**	**100**	**250**	**500**	**1000**	**2500**	**100000**
SEM ML rel bias	-0.205	-0.175	-0.166	-0.168	-0.161	-0.166	-0.164
SEM ULS rel bias	-0.139	-0.145	-0.159	-0.167	-0.163	-0.170	-0.169
LSAM ML rel bias	-0.498	-0.385	-0.272	-0.225	-0.196	-0.189	-0.180
LSAM ULS rel bias	-0.497	-0.385	-0.272	-0.225	-0.196	-0.189	-0.180
GSAM ML rel bias	-0.497	-0.385	-0.272	-0.225	-0.196	-0.189	-0.180
GSAM ULS rel bias	-0.496	-0.385	-0.272	-0.225	-0.196	-0.189	-0.180

Note. SEM ML = Maximum-likelihood estimation, SEM ULS = Unweighted-least-squares estimation, LSAM ML = Local-SAM-maximum-likelihood estimation, LSAM ULS = Local-SAM-unweighted-least-squares estimation, GSAM ML = Global-SAM-maximum-likelihood estimation, GSAM ULS = Global-SAM-unweighted-least-squares estimation.

Table 5

Study 2 (Kosanke): RMSE in conditions with two negative unmodelled cross-loadings.

	Sample Size						
Method/Metric	**50**	**100**	**250**	**500**	**1000**	**2500**	**100000**
SEM ML	0.48	0.382	0.257	0.211	0.182	0.172	0.161
SEM ULS	0.477	0.367	0.241	0.205	0.182	0.175	0.166
LSAM ML	0.486	0.421	0.323	0.269	0.232	0.216	0.201
LSAM ULS	0.487	0.421	0.323	0.269	0.232	0.216	0.201
GSAM ML	0.49	0.422	0.323	0.269	0.232	0.216	0.201
GSAM ULS	0.49	0.421	0.323	0.269	0.232	0.216	0.201

Note. SEM ML = Maximum-likelihood estimation, SEM ULS = Unweighted-least-squares estimation, LSAM ML = Local-SAM-maximum-likelihood estimation, LSAM ULS = Local-SAM-unweighted-least-squares estimation, GSAM ML = Global-SAM-maximum-likelihood estimation, GSAM ULS = Global-SAM-unweighted-least-squares estimation.

Table 6

Study 3 (Kosanke): Relative bias in conditions with each one positive unmodelled cross-loading and residual correlation.

	Sample Size						
Method/Metric	**50**	**100**	**250**	**500**	**1000**	**2500**	**100000**
SEM ML rel bias	0.209	0.270	0.283	0.289	0.289	0.282	0.284
SEM ULS rel bias	0.250	0.284	0.277	0.280	0.279	0.271	0.272
LSAM ML rel bias	-0.232	-0.061	0.127	0.211	0.246	0.261	0.276
LSAM ULS rel bias	-0.229	-0.060	0.127	0.211	0.246	0.261	0.276
GSAM ML rel bias	-0.230	-0.060	0.127	0.211	0.246	0.261	0.276
GSAM ULS rel bias	-0.228	-0.060	0.127	0.211	0.246	0.261	0.276

Note. SEM ML = Maximum-likelihood estimation, SEM ULS = Unweighted-least-squares estimation, LSAM ML = Local-SAM-maximum-likelihood estimation, LSAM ULS = Local-SAM-unweighted-least-squares estimation, GSAM ML = Global-SAM-maximum-likelihood estimation, GSAM ULS = Global-SAM-unweighted-least-squares estimation.

In Study 4, a comparative advantage of LSAM compared to SEM-ML was present, but only for smaller samples. With regards to RMSE, results were mixed as well. LSAM appeared to outperform in Table 7 for DGM 2 of Study 4. In other conditions, however, no substantial

differences arose in terms of RMSE.

Small sample bias in LSAM estimation *The small sample bias of LSAM estimation in Study 1b revealed that in smaller samples ranging from N=50 to N=100, both LSAM-ML and LSAM-ULS estimation were biased. Table 8 shows this was especially apparent in a sample size of 50.*

Table 7

Study 4 (Kosanke): RMSE in DGM 2 (conditions with five positive unmodelled cross-loadings).

	Sample Size						
Method/Metric	**50**	**100**	**250**	**500**	**1000**	**2500**	**100000**
SEM ML	0.373	0.257	0.166	0.124	0.107	0.100	0.095
SEM ULS	2.070	0.373	0.320	0.306	0.300	0.298	0.296
LSAM ML	0.188	0.141	0.103	0.089	0.080	0.075	0.071

Note. SEM ML = Maximum-likelihood estimation, SEM ULS = Unweighted-least-squares estimation, LSAM ML = Local-SAM-maximum-likelihood estimation.

Table 8

Study 1b (Kosanke): Absolute bias of LSAM-ML for N=50.

	Phi Levels				
Lambda	**0**	**0.2**	**0.4**	**0.6**	**0.8**
0.4	0.202	0.202	0.258	0.346	0.444
0.5	0.187	0.179	0.203	0.245	0.305
0.6	0.176	0.165	0.166	0.170	0.190
0.7	0.164	0.150	0.139	0.122	0.116
0.8	0.150	0.135	0.119	0.098	0.074

Note. LSAM ML = Local-SAM-maximum-likelihood estimation.

Note that the absolute values of bias were calculated in this study. Consequently, the values of the bias should be interpreted as negative, as follows from the results of the original paper (Robitzsch, 2022). The bias persisted, but to a lesser degree in samples of 100. Thus, as expected, a clear effect of sample size was present. Overall, comparing LSAM-ML and -ULS estimation, results were very similar. Importantly, differential effects due to lambda and phi were present in Study 1b. The small sample bias was especially strong for lower lambda and higher phi values, thus in contexts of low reliability and high factor correlations. Also, a new insight is that

the bias remained relevant for low values of phi, unlike in the original paper by Robitzsch (2022). In consequence, there seemed to be no conditions were SAM's small sample bias was negligible. Another new insight lied in the presence of what could be called a reversal effect: For higher values of lambda, the bias did not increase for higher values of phi. On the contrary, absolute bias values decreased for higher phi values, when looking at the conditions with lambda = 0.7-0.8. As an additional investigation of the small sample bias, I included Study 4a to see its effect come to play in a 5-factor-model with regressions. Table 9 shows the performance of SEM-ML, whereas Table 10 shows the performance of LSAM-ML in DGM 2 (in presence of unmodelled cross-loadings). Aligning with the findings of Study 1b, the results suggested an even better relative performance of LSAM- over traditional SEM-ML estimation for smaller N and higher beta. Thus, the negative small sample bias came into play in this study as well. Results looked very similar with regards to RMSE. Note that this trend was less strong, but still present in the conditions of DGM 3 when looking at residual correlations.

Table 9

Study 4a (Kosanke): Absolute bias of SEM-ML for DGM 2 (conditions with five positive unmodelled cross-loadings).

	Sample Size						
Beta	**50**	**100**	**250**	**500**	**1000**	**2500**	**100000**
0.1	0.253	0.172	0.115	0.095	0.085	0.080	0.075
0.2	0.327	0.220	0.162	0.139	0.122	0.114	0.109
0.3	0.545	0.270	0.218	0.206	0.195	0.190	0.180
0.4	0.981	0.336	0.270	0.257	0.252	0.251	0.254

Note. SEM ML = Maximum-likelihood estimation, SEM ULS = Unweighted-least-squares estimation, LSAM ML = Local-SAM-maximum-likelihood estimation, LSAM ULS = Local-SAM-unweighted-least-squares estimation, GSAM ML = Global-SAM-maximum-likelihood estimation, GSAM ULS = Global-SAM-unweighted-least-squares estimation.

Table 10

Study 4a (Kosanke): Absolute bias of LSAM-ML for DGM 2 (conditions with five positive unmodelled cross-loadings).

	Sample Size						
Beta	**50**	**100**	**250**	**500**	**1000**	**2500**	**100000**
0.1	0.150	0.112	0.083	0.072	0.065	0.061	0.057
0.2	0.145	0.108	0.077	0.066	0.059	0.056	0.053

Table 10

Study 4a (Kosanke): Absolute bias of LSAM-ML for DGM 2 (conditions with five positive unmodelled cross-loadings). (continued)

Beta	50	100	250	500	1000	2500	100000
0.3	0.141	0.104	0.073	0.060	0.052	0.047	0.043
0.4	0.151	0.115	0.088	0.076	0.070	0.067	0.065

One aspect to mention is that lambda values were quite high in study 4a (lambda=0.7). Matching the results from Study 1b, SAM's bias did not increase for higher values of beta, unlike SEM's. This effect could hint at a stronger robustness of SAM in contexts of higher correlations with misspecifications present. But, as the effect could not be observed in other conditions (e.g. in DGM 3 with residual correlations), I did not deem it substantial.

Summary of results *For the most part, I succesfully replicated the results from Robitzsch (2022): I did not observe substantial convergence issues in any study. Across studies, as in the original paper, SAM did not generally outperform SEM in small to moderate samples. SAM exhibited a negative small sample bias that made SAM appear superior in conditions with unmodelled positive cross-loadings and residual correlations. This bias was especially strong for lower lambda and higher phi or beta values. Going ahead of what was investigated in Robitzsch (2022), I found that this bias is also present in models with lower phi or beta values. Thus, it cannot be concluded that SAM is more robust in models with non-saturated structural parameters. If there was no misspecification or unmodelled negative cross-loadings and residual correlations, SAM tended to perform worse than traditional SEM, as far as can be concluded from my results.*

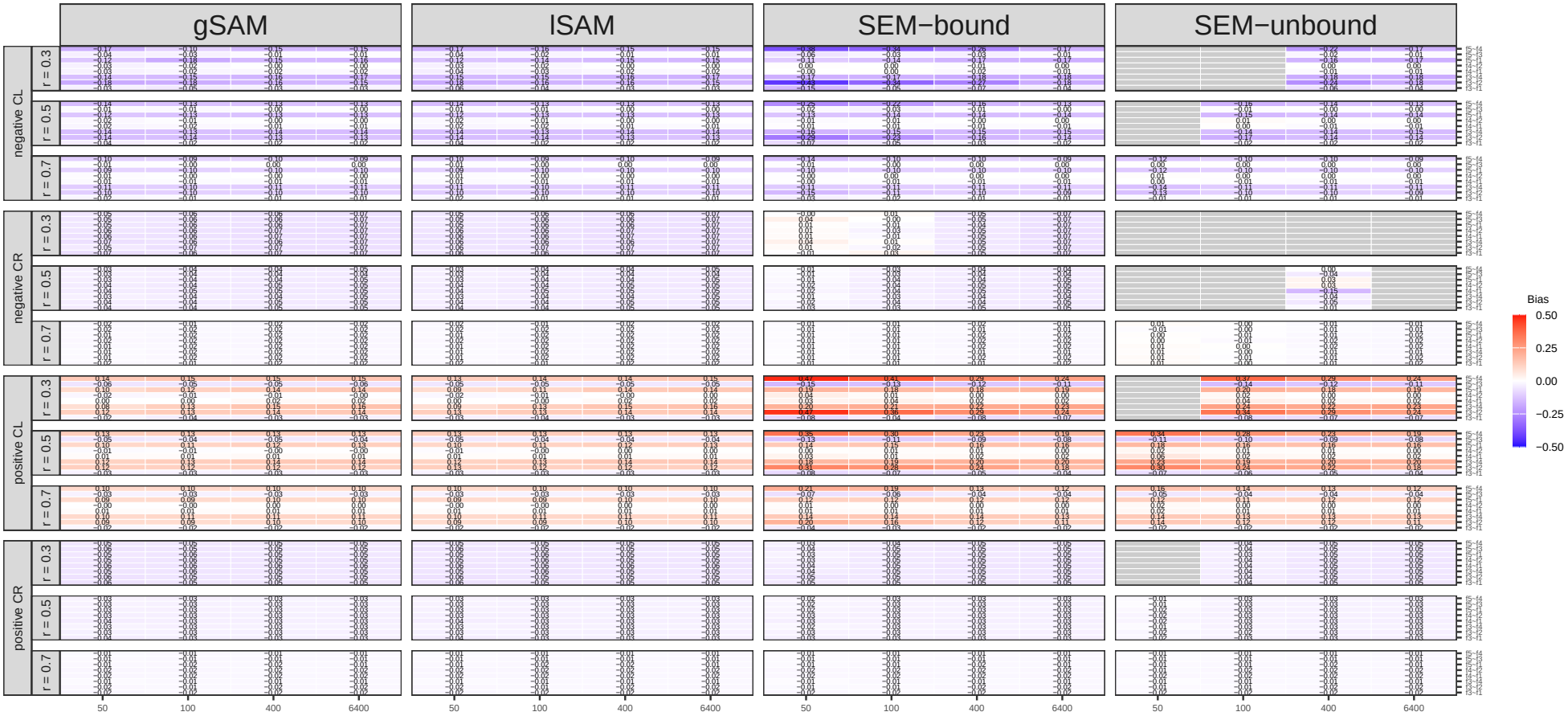
Combined Simulation Study

The combined study showed that using bound maximum likelihood estimation for standard SEM as proposed by Collaborator B did, in fact, eliminate the low convergence rate as well as improper solutions in all conditions (see B1). Next, the parameter-wise signed mean bias values of the single regression weight estimates showed that omitting positive cross-loadings results in an overall positive bias, whereas omitting negative cross-loadings results in a negative bias. As in previous studies, the positive bias was less pronounced for SAM (gSAM and lSAM) than for SEM, especially with lower sample sizes and reliability. However, contrary to the findings by Collaborator B and Robitzsch (2022), in this study, the negative bias was also less pronounced for both SAM methods. Conditions with omitted correlated residual correlations resulted in predominantly negative bias values for all methods irrespective of the sign of the misspecification,

with standard SEM being slightly less biased than SAM methods (see Figure 9). A direct comparison of standard SEM with SAM for bias and RMSE showed that SAM methods were less biased for cross-loadings and, when accounting for variance via RMSE, more accurate, especially in smaller sample sizes and indicator reliability levels (see Figure 10). Further, there were little to no differences between gSAM and lSAM, with lSAM being slightly more accurate overall in terms of RMSE. Overall these findings suggest that the explanation for SAM's lower bias under positive misspecifications being due to a general negative bias that would lead to overly negative bias in case of negative measurement misspecifications (proposed by Collaborator B and Robitzsch (2022)) does not hold in a more complex (and realistic) model that SAM could be favorable choice of SEM estimation especially challenging conditions.

Figure 9

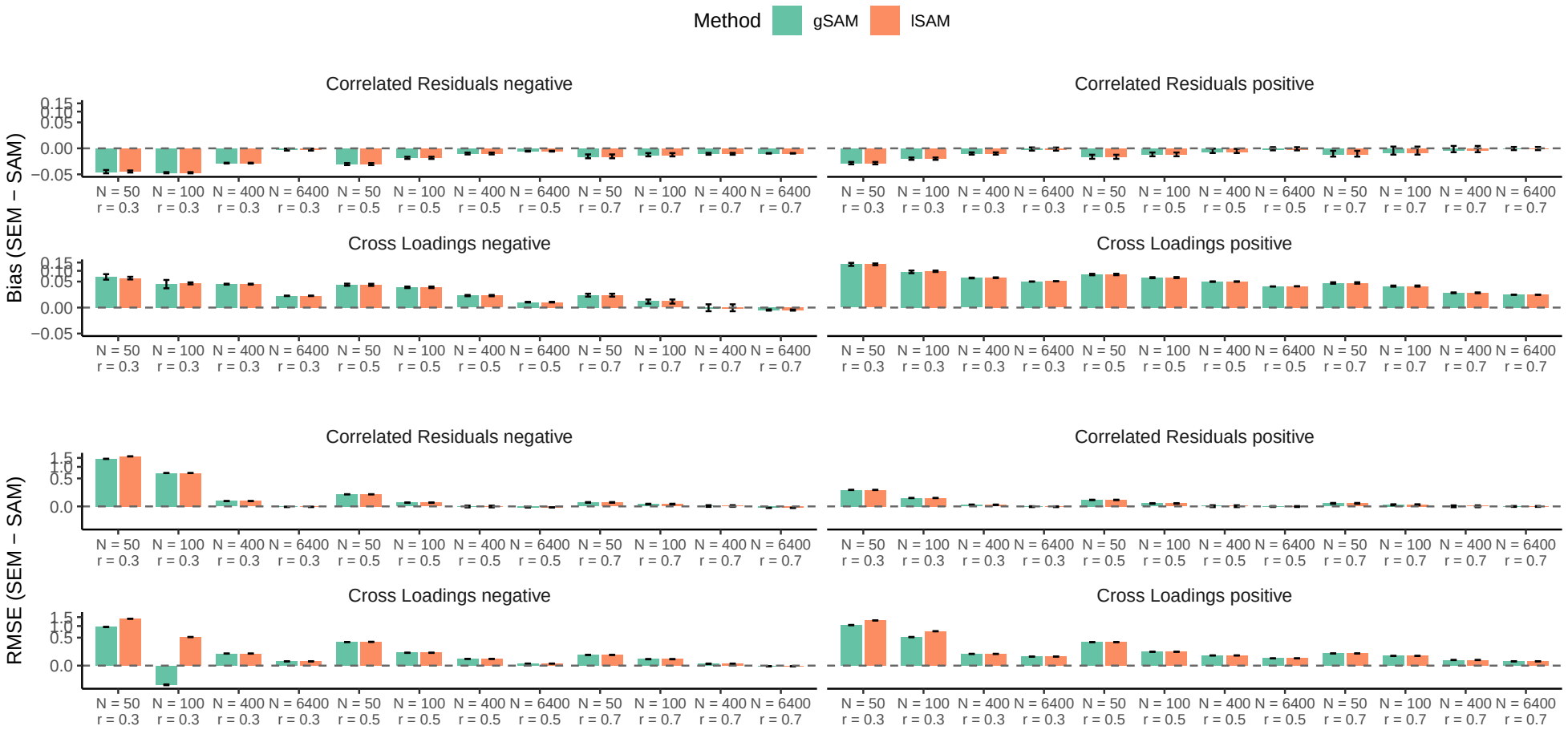
Mean Bias of Regression Parameters in Joint *Study*



Note. Mean absolute bias for each parameter with a true value of 0.1 in one model for sample sizes (N), reliability (r), and misspecifications for global SAM (gSAM), local SAM with Maximum Likelihood (lSAM-ML), Unweighted Least Squares (lSAM-ULS) and SEM. Grey fields indicate excluded conditions with convergence rate below 50%. The Figure is adapted from Dhaene & Rosseel (2023).

Figure 10

Aggregated Mean Difference between Bound Standard SEM and SAM Estimation in Bias and RMSE of regression coefficients estimates in Joint *Study*



Note. bias (top) and RMSE (bottom) differences between SEM and global and local SAM (gSAM and lSAM), averaged across estimates of true regression coefficients (0.1) over varying sample size (N), reliability (r), and misspecifications. Error bars indicate Monte Carlo Standard Errors.

Adversarial Collaboration Process

Collaborator B (Kosanke) chose to terminate the AC after the initial stage of individual studies. He argued that while the Structural After Measurement (SAM) approach generally showed less bias and root mean square error (RMSE), there were certain conditions — particularly those involving negative unmodeled residuals and cross-loadings — where the advantages of traditional Structural Equation Modeling (SEM) offset those of SAM. This finding aligns with Robitzsch (2022), who observed a general negative bias in SAM. Further, he reasoned that Collaborator A's findings did not directly contradict this result, and the integration of the individual studies supported a general conclusion: SAM does not consistently outperform SEM in small to moderate samples. However, imporantly, Collaborator B acknowledged that his findings were limited to simple two to three-factor models without directed strucutral paths and that further more targeted investigation in more complex models was warranted and conceded that his prdominant reason for the termination of the AC was due to time constraints.

Thus, following this, Collaborator A (Kriegmair) identified multiple arguments for conducting an additional simulation based on the initial round of replicated studies and proceeded to establish an integrated study. Collaborator B's conclusion regarding SAM's inconsistent outperformance of SEM in the presence of negative misspecifications was drawn only from a specific type of model with no directed relationships between factors and had not been tested in a more complex model with directed structural paths of interest. Importantly, only the latter represents the type of model for which Rosseel & Loh (2022) suggested SAM to be advantageous. Moreover, to thoroughly investigate the hypothesized systematic underestimation by SAM, a parameter-wise analysis of bias was warranted. Aggregating bias values across model parameters could mask effects by averaging out negative and positive values or obscuring them entirely when using absolute values. Further, the claim that constrained SEM estimation methods could mitigate convergence issues and improper solutions was not thoroughly tested in the individual studies. Although the collaborators did not jointly conclude to conduct a collaborative, unified simulation study as initially planned, the individual studies provided a solid foundation for a joint study. In this context, a unified study was conducted to formally complete the AC process, involving a stepwise integration of settings from the individual studies to address conflicting verbal claims as presented in the following while full details simulation details are available in the methods section and in the full simulation protocol in Appendix A.

Population Model and Data Generating Mechanism In the combined study for parametric data generation, a five-factor population structural model with three indicators per factor, based on Collaborator A's approach, was selected to provide a complex model suitable for testing the advantages of SAM. CFA models from Collaborator B's studies were excluded, as

SAM is intended for models with directed structural paths. Indicator reliability was manipulated by adjusting measurement error variances, following Collaborator A's method, to achieve a valid representation of item reliability.

Experimental Design To evaluate the robustness of SAM, especially in response to negative misspecifications identified by Collaborator B, both positive and negative omitted cross-loadings and correlated residuals were included in the model misspecifications. The experimental design extended sample size variation to include very small samples (N = 50), providing a more comprehensive range than in Collaborator A's studies but very large (population level) samples (N = 10). To limit the scope and narrow down on conditions relevant to expected differences between SAM and standard SEM, varying number of measurement blocks for SAM and strength of structural relations was not carried over from Collaborator A's studies as the results did not suggest a substantial impact on SAM compared to SEM.

Method Selection In selecting estimation methods, bound SEM-ML, unbound SEM-ML, gSAM, and lSAM-ML were compared to address convergence issues with standard SEM, as proposed by Collaborator B. SAM-ULS and SEM-ULS were excluded to focus on the most relevant methods and manage computational demands.

Estimands and Performance Measures As metrics bias and RMSE of the estimated structural parameters were calculated, and parameter-wise bias analysis was conducted to identify any potential negative bias in SAM, eliminating the effects of averaging.

Analysis The analysis was broadly consistent with the individual studies, with the addition of parameter-wise bias displays to provide deeper insights into the biases and direct computation of metric differences between SEM and SAM.

Discussion

The goal of the current study was twofold: First, to test the viability and practical applicability of Adversarial Collaboration (AC) as a tool to resolve disagreeing research claims and enhance generalizability and rigor in the context of simulation studies. Second, serving as a case study for this first aim, to evaluate the performance of traditional Structural Equation Modeling (SEM) compared to Structural After Measurement (SAM) and resolve conflicting claims of previous studies regarding whether SAM consistently outperforms traditional SEM in the presence of model misspecifications in small to moderate sample sizes (Dhaene & Rosseel, 2023; Robitzsch, 2022; Rosseel & Loh, 2022).

SAM vs. SEM

Overall, this case study systematically assessed how different implementations of SAM and standard SEM performed under varying sample sizes, indicator reliabilities, and degrees of model misspecification. In Collaborator A's (Kriegmair) individual studies, SAM methods—including global SAM (gSAM) and local SAM (lSAM) with Maximum Likelihood (ML) and Unweighted Least Squares (ULS) estimators—consistently outperformed traditional SEM, particularly under challenging conditions that replicated previous findings (Dhaene & Rosseel, 2023; Rosseel & Loh, 2022). Under these conditions, SAM methods achieved higher convergence rates and demonstrated lower average relative biases and RMSE values compared to standard, unconstrained SEM. Standard SEM exhibited significant convergence issues and a higher incidence of improper solutions, especially when models were misspecified by omitting cross-loadings or correlated residuals. Collaborator B's (Kosanke) individual studies replicated findings from Robitzsch (2022), indicating that SAM did not generally outperform SEM in small to moderate samples. According to this, SAM exhibited a negative small-sample bias, which made it appear superior only under conditions with unmodeled positive cross-loadings and residual correlations. Previous research has shown that omitting existing cross-loadings or correlated residuals in SEM measurement models, effectively constraining them to zero, can result in absorption of the unmodelled variance by the latent factor relations in the structural model of the same direction as the unmodelled parameters (Ferrando et al., 2022; Hsu et al., 2014; Lance et al., 2010; Steenkamp & Maydeu-Olivares, 2023). Thus, a general underestimation of structural parameters by SAM was argued to offset the positive bias introduced by the unmodeled positive cross-loadings and residual correlations in Collaborator A's studies, creating an appearance of greater accuracy compared to SEM. In conditions without misspecification or with unmodeled negative cross-loadings and residual correlations, SEM tended to outperform SAM. However, these conclusions mainly rely on simple two to three-factor models without directed latent relations of interest. The combined study sought to reconcile these differing findings in more complex models representative of use cases of SAM. First, by applying bound maximum likelihood estimation for SEM, as proposed by Kosanke, the convergence issues and improper solutions previously observed in SEM were effectively eliminated across all conditions, even in complex five-factor models with an extensive structural model. Second, parameter-wise analysis revealed that omitting positive cross-loadings led to an overall positive bias, while omitting negative cross-loadings led to a negative bias, as was previously shown (Hsu et al., 2014; Steenkamp & Maydeu-Olivares, 2023). Notably, SAM methods exhibited less pronounced biases in both cases compared to SEM, especially at lower sample sizes and reliability levels. Contrary to Collaborator B's findings, the negative bias was also less pronounced for SAM methods in the

combined study. When directly comparing SEM and SAM in terms of bias and RMSE, SAM methods were generally less biased and more accurate, particularly in conditions involving omitted cross-loadings and in scenarios with smaller sample sizes and lower indicator reliability. Moreover, in contrast to previous findings in almost all conditions, all methods showed systematic under- rather than over-estimation of parameters in the presence of positive (as well as negative) misspecifications (Ferrando et al., 2022). Finally, differences between gSAM and lSAM were minimal, with lSAM showing a slight advantage in RMSE. Additionally, the current results support constrained standard SEM as a promising alternative for complex models with relatively low sample sizes, given its improved convergence and stability compared to unconstrained standard SEM, with the benefits in convergence possibly outweighing the potential risk of obscuring model misfit (Savalei & Kolenikov, 2008).

Collectively, these findings suggest that SAM methods may offer advantages over traditional SEM in handling model misspecifications, particularly under challenging conditions. While the individual studies, replicating Robitzsch (2022), highlighted potential biases associated with SAM in small samples, the combined study indicates that these biases are not sustained in complex models. This supports the notion that SAM could be a favourable choice for SEM estimation in practice, particularly when addressing small sample sizes, low reliability, and potential model misspecifications, especially in more complex models.

However, the simulations conducted in this study have several limitations. They focused on a limited subset of models and conditions, constraining the generalizability of the findings to specific scenarios where the analyzed methods exhibit differential performance. The measurement models for most simulated data were restricted to three indicators per factor, without variation across factors. Apart from certain models where factor loadings were adjusted to influence reliability, these loadings were generally kept constant and unrealistically high (between 0.7 and 0.9), following the studies by Rosseel & Loh (2022) and Dhaene & Rosseel (2023). Additional modulation of factor loadings, alongside reliability modulation via measurement error variances, could have provided more realistic and extensive simulated data. Future research could explore the performance of Structural After Measurement (SAM) in longitudinal data contexts, particularly with causal models, as SAM is aimed at applications where structural parameters are central, and measurement models may be (potentially) misspecified (Rosseel & Loh, 2022). Additionally, it would be valuable to assess the differential impact of SAM on various structural parameters, such as mediating versus direct effects within mediation models. As suggested by Rosseel & Loh (2022), SAM is particularly advantageous for scenarios involving small sample sizes relative to model complexity. Future studies could include conditions that systematically vary model complexity alongside sample size, potentially leading to the development of a metric to balance model complexity and sample size.

Adversarial Collaboration in Simulation Studies

Shifting perspective, in our collaborative effort, we successfully established a joint starting point by translating conflicting verbal claims from prior studies into shared research questions. Based on these questions, we independently conducted simulation studies, largely drawing on the methodologies of Rosseel & Loh (2022), Dhaene & Rosseel (2023), and Robitzsch (2022). This approach effectively translated the verbal dispute into an empirical investigation. Notably, this process was only an emulation of AC, incorporating an additional layer by replicating previously published research findings in the first round of our framework. In a practical application of AC to simulation studies, as proposed here, this intermediate step could be omitted. Instead, collaborators might design two original studies or opt to work directly together on a unified research study, depending on the identification of a specific verbal disagreement. Furthermore, although the unified simulation study was conducted by one collaborator, the individual studies highlighted both areas of agreement and divergence. While Collaborator B concluded that neither SAM nor SEM consistently outperformed the other across broader applications, The need for further exploration of SAM's performance in more complex models with directed structural paths und negative misspecifications was identified and acknowledged by both researchers. Additionally, even if conducted only by one collaborator, the unified study serves as a proof of concept regarding the techincal fesability for applying AC to simulation studies demonstrating how an actual collaboration process could unfold via a joint study.

In conclusion, the current study demonstrates that AC is a practically applicable and viable approach to further scientific rigor and potentially increase generalizability in the context of Monte Carlo simulation studies. By successfully translating a general conflict in findings about the performance of SAM and standard SEM into a joint research question and, based on this, directly juxtaposing our different simulation setups, we were able to trace back general diverging conclusions to specific methodological operationalizations and technical decisions. This effectively enhanced transparency and reduced ambiguity of the conflicting claims, allowing for more precise identification of the sources of disagreement, akin to previous implementations of AC in empirical research (Mellers et al., 2001; Melloni et al., 2023). In particular, this approach enforced direct engagement with each adversary's specific arguments for their conclusions, transparently linking each claim to its operationalized source and integrating it into a combined study aimed at falsifying the conflicting claims. This demonstrates how the practice of AC can aid in moving simulation studies closer to a Popperian ideal of falsificationism (Lakatos et al., 1978; Popper, 1963) by systematically testing and potentially refuting conflicting claims in a structured and transparent way instead of designing explicitly or implicitly confirmationist studies tailored to corroborating specific claims about a method's performance. Hence, adversarial simulation

appears as a promising tool to enhance generalizability and rigor beyond other open science practices (O'Kelly et al., 2017; Pawel et al., 2023) at the point of simulation design without overly constraining researchers degrees of freedom in hypothesis generation and operationalization (Buchka et al., 2021; Clark et al., 2022a). Notably, this falsificationistic nature of AC should not be misunderstood as a strict adherence to a binary logic of true or false claims; it has also been suggested to align with a Bayesian framework of belief updating and evidence accumulation, serving as a systematic and transparent method to reduce ambiguity and increase generalizability by testing conflicting claims in a structured way (Corcoran et al., 2023). In the current study AC led to more targeted, less ambiguous, and arguably more generalizable results and conclusions by adjusting multiple methodological aspects, such as model type, misspecifications, reliability computation, sample size, and analysis. However, it is important to note that due to the specific circumstances of this case study, which included a prior replication of previous results, these results impacted the collaboration and the combined study. If this step were omitted in a practical application of AC, the combined study would potentially need to be more comprehensive, covering a broader range of settings, if prior results would not be as readily available to inform it. Nevertheless, this initial independent replication phase has the advantage of isolating initial discrepancies for preexisting disagreements, which can clarify specific origins of divergences. As recently demonstrated in an empirical setting by Melloni et al. (2023) involving cross-lab replications in an AC project, such a step can help identify key factors that contribute to preexisting conflicting claims, allowing collaborators to later design a more streamlined and focused joint study. This connects to another possible objection: one could argue that most cases of contrasting simulation studies could be resolved by just extending the simulation to include all relevant settings, and this objection is partly valid, as demonstrated, for example, by the extension of the sample size factor in the combined study. However, as also shown by this study, there are diverging operationalizations that are exclusively and directly tied to specific conclusions, which are unresolvable by extending the simulation settings. Furthermore, individual authors might be explicitly or implicitly tempted toward biased conclusions from overly extensive simulation settings when mapping the results back to the general level, cherry-picking the most favourable results and clouding a clear and transparent interpretation (Buchka et al., 2021; Clark et al., 2022a). Finally, the current study also demonstrates how Adversarial Collaboration can be advantagously combined with other open science practices, such as preregistration, open source, open data and reproducability practices. Here the preregistration of the individual studies, the availabitilty of open source code and data on GitHub as well as cross reproducability via containerization

Albeit promising, *Adversarial Simulation* also presented several challenges and limitations. First, as already emphasized by Clark et al. (2022b), AC demands increased resources in terms of

time and coordination. The process requires careful planning, open communication, and a willingness to reconcile differing viewpoints, and it may only justify its high cost if it advances science by rigorously clarifying core issues in disputes and not as a default. In our case, the termination of the collaboration by one party underscores the potential difficulties in sustaining such efforts. Even though simulation studies are less resource-intensive than empirical data acquisition, here, too, the additional relative time and effort required may pose practical constraints, especially in academic environments with tight schedules and resource limitations. Second, even after collaborating and merging, the individual studies and settings of the combined study remained limited to very specific conditions, and any general conclusions from its results still rely on induction with the premise of having considered a prototypical scenario representing various application settings (Bollmann et al., 2015; Feinberg & Rubright, 2016; Gilbert & Miratrix, 2024). Third, the current project constitutes only an artificial case of AC, where the collaborators had no disagreement based on personal prior research but were assigned to represent conflicting viewpoints without any stakes or personal investment in the outcome. Furthermore, the final stage of collaboratively designing and conducting a study was conducted by one collaborator. Thus, this merely serves as a proof of concept showcasing the potential and technical applicability of AC in simulation studies. Fourth, the applicability of AC may be limited to specific settings where there are clear, conflicting viewpoints on particular methods or theories. It may not be as effective in areas where disagreements are less defined or more complex. The focus on specific methodological disputes means that broader issues or more subtle disagreements might not be as amenable to this approach (Clark & Tetlock, 2023) However, it can align with the recently recognized need for rigorous and unbiased assessment of comparative simulation studies in methodological research (Boulesteix et al., 2013) by providing an approach for researchers to juxtapose different (and possibly their own) statistical methods against each other. Despite these challenges, overall, this case study shows that AC has the potential to increase generalizability and rigor in simulation studies beyond other open science practices by promoting transparency and reducing biases. By bringing together researchers with opposing views, it encourages a more thorough examination of assumptions and methodological choices, potentially leading to more robust and reliable conclusions.

An alternative approach to address the generalizability challenge, specifically at the point of operationalization and simulation design, is to ground simulations in empirical data. By incorporating actual models and parameters from practice (Bollmann et al., 2015), sampling models and parameters from the literature can effectively bridge the gap between prototypical simulations and real-world applications, more directly targeting the generalizability challenge of simulation studies. Another avenue for future research inspired by this project is the development of collaborative simulation platforms that could facilitate ongoing contributions from multiple

researchers. By leveraging open-source tools and platforms such as GitHub, similar to the current project but more elaborate and refined, simulations can be made "living" open-source projects that are continuously updated and refined. Collaborators could utilize open-source platforms like GitHub to add new conditions or settings to existing simulations, allowing for dynamic testing and continuous integration of new results. For example, a research team interested in assessing the performance of one method for a specific type of model and data relevant to their work could contribute to an existing simulation repository. With minimal coding effort, they could add their conditions, request a rerun of the simulation, and obtain updated results that inform both their specific research and the general understanding of the method's performance. This collaborative approach could effectively enhance the generalizability of simulation studies, making them more accessible and responsive to the needs of the research community and better capturing the complexities across different contexts. Implementing such collaborative simulation platforms would require computational infrastructure and a system for regulating when simulations are rerun to manage computational costs. Streamlined pipelines involving containerization technologies like Docker or Singularity (Kurtzer et al., 2017; Merkel, 2014) could facilitate the dynamic deployment of simulations on high-performance computing resources. Additionally, establishing guidelines and open peer review mechanisms for contributions would ensure the quality and integrity of the simulations. Finally, the comparison of the individual studies and the process of conducting a combined study also highlighted the need for a more structured and standardized code of conduct for programming the simulation studies that could facilitate collaboration. This could include a common modular file structure and naming conventions, standardized code comments, and documentation.

Scientific progress relies on a delicate balance between general theoretical claims, concrete operationalizations, and rigorous testing where researchers degrees of freedom are essential for designing and conducting innovative and impactful research. However, these same freedoms can introduce ambiguity and bias, posing risks to the reproducibility and generalizability of findings. In sum this thesis demonstrates how Adversarial Collaboration can effectively map onto this dynamic in the context of Monte Carlo simulation studies, offering a flexible and yet rigorous testing ground where conflicting general claims can be systematically examined and refined through operationalized inquiry, embracing (methodological) research as a communal endeavor.

References

Banack, H. R., Hayes-Larson, E., & Mayeda, E. R. (2021). Monte carlo simulation approaches for quantitative bias analysis: A tutorial. *Epidemiologic Reviews*, *43*(1), 106–117. https://doi.org/10.1093/epirev/mxab012

Bartlett, M. S. (1937). The Statistical Conception of Mental Factors. *British Journal of Psychology. General Section*, *28*(1), 97–104. https://www.proquest.com/docview/1293463650/citation/9FE530638F5041E0PQ/1

Bartlett, M. S. (1938). Methods of estimating mental factors. *Nature*.

Bollen, K. A. (2014). *Structural equations with latent variables*. John Wiley & Sons.

Bollmann, S., Heene, M., Küchenhoff, H., & Bühner, M. (2015). *What can the Real World do for simulation studies? A comparison of exploratory methods*. https://doi.org/10.5282/UBM/EPUB.24518

Boulesteix, A.-L., Lauer, S., & Eugster, M. J. A. (2013). A Plea for Neutral Comparison Studies in Computational Sciences. *PLoS ONE*, *8*(4), e61562. https://doi.org/10.1371/journal.pone.0061562

Buchka, S., Hapfelmeier, A., Gardner, P. P., Wilson, R., & Boulesteix, A.-L. (2021). On the optimistic performance evaluation of newly introduced bioinformatic methods. *Genome Biology*, *22*(1), 152. https://doi.org/10.1186/s13059-021-02365-4

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., et al. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644.

Carrillo, N., & Martínez, S. (2023). Scientific inquiry: From metaphors to abstraction. *Perspectives on Science*, *31*(2), 233–261. https://doi.org/10.1162/posc_a_00571

Clark, C. J., Costello, T., Mitchell, G., & Tetlock, P. E. (2022a). Keep your enemies close: Adversarial collaborations will improve behavioral science. *Journal of Applied Research in Memory and Cognition*, *11*(1), 1–18. https://doi.org/10.1037/mac0000004

Clark, C. J., Costello, T., Mitchell, G., & Tetlock, P. E. (2022b). The road less traveled: Understanding adversaries is hard but smarter than ignoring them. *Journal of Applied Research in Memory and Cognition*, *11*(1), 50–53. https://doi.org/10.1037/mac0000020

Clark, C. J., & Tetlock, P. E. (2023). Adversarial collaboration: The next science reform. In *Ideological and political bias in psychology: Nature, scope, and solutions* (pp. 905–927). Springer.

Clark, C., & Tetlock, P. (2021). *Adversarial collaboration: The next science reform*. https://doi.org/10.1007/978-3-031-29148-7_32

Cleeremans, A. (2022). Theory as adversarial collaboration. *Nature Human Behaviour*, *6*(4), 485–486.

Corcoran, A. W., Hohwy, J., & Friston, K. J. (2023). Accelerating scientific progress through bayesian adversarial collaboration. *Neuron*, *111*(22), 3505–3516.

Dellsén, F., & Baghramian, M. (2021). Disagreement in science: Introduction to the special issue. *Synthese*, *198*(Suppl 25), 6011–6021.

Dhaene, S., & Rosseel, Y. (2023). An Evaluation of Non-Iterative Estimators in the Structural after Measurement (SAM) Approach to Structural Equation Modeling (SEM). *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(6), 926–940. https://doi.org/10.1080/10705511.2023.2220135

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00621

Feinberg, R. A., & Rubright, J. D. (2016). Conducting Simulation Studies in Psychometrics. *Educational Measurement: Issues and Practice*, *35*(2), 36–49. https://doi.org/10.1111/emip.12111

Ferrando, P. J., Hernandez-Dorado, A., & Lorenzo-Seva, U. (2022). Detecting correlated residuals in exploratory factor analysis: New proposals and a comparison of procedures. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(4), 630–638.

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465.

Gilbert, J., & Miratrix, L. (2024). Multilevel metamodels: A novel approach to enhance efficiency and generalizability in monte carlo simulation studies. *arXiv Preprint arXiv:2401.07294*.

Hair Jr, J. F., Hult, G. T. M., Ringle, C. M., Sarstedt, M., Danks, N. P., Ray, S., Hair, J. F., Hult, G. T. M., Ringle, C. M., Sarstedt, M., et al. (2021). An introduction to structural equation modeling. *Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R: A Workbook*, 1–29.

Hamaker, E. L. (2023). The within-between dispute in cross-lagged panel research and how to move forward. *Psychological Methods*.

Hoyle, R. H. (2012). *Handbook of structural equation modeling*. Guilford press.

Hsu, H.-Y., Skidmore, S. T., Li, Y., & Thompson, B. (2014). Forced zero cross-loading misspecifications in measurement component of structural equation models. *Methodology*.

Joshi, M., & Pustejovsky, J. (2024). *Simhelpers: Helper functions for simulation studies*. https://meghapsimatrix.github.io/simhelpers

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515.

Kline, R. B. (2023). *Principles and practice of structural equation modeling*. Guilford publications.

Kriegmair, V. (2024). *Preregistration: Comparing a structural after measurement (SAM) approach to standard structural equation model (SEM) estimation*. Zenodo. https://doi.org/10.5281/zenodo.11459378

Krijnen, W. P. (1996). Algorithms for unweighted least-squares factor analysis. *Computational Statistics & Data Analysis*, *21*(2), 133–147.

Kulinskaya, E., Hoaglin, D. C., & Bakbergenuly, I. (2020). *Exploring Consequences of Simulation Design for Apparent Performance of Statistical Methods. 1: Results from simulations with constant sample sizes*. arXiv. http://arxiv.org/abs/2006.16638

Kurtzer, G. M., Sochat, V., & Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. *PloS One*, *12*(5), e0177459.

Lakatos, I. et al. (1978). *The methodology of scientific research programmes*. Cambridge: Cambridge university press.

Lance, C. E., Lance, C. E., & Vandenberg, R. J. (2010). On the practice of allowing correlated residuals among indicators in structural equation models. In *Statistical and methodological myths and urban legends* (pp. 213–236). Routledge.

Latham, G. P., Erez, M., & Locke, E. A. (1988). Resolving scientific disputes by the joint design of crucial experiments by the antagonists: Application to the erez–latham dispute regarding participation in goal setting. *Journal of Applied Psychology*, *73*(4), 753.

Lüdtke, O., Ulitzsch, E., & Robitzsch, A. (2021). A Comparison of Penalized Maximum Likelihood Estimation and Markov Chain Monte Carlo Techniques for Estimating Confirmatory Factor Analysis Models With Small Sample Sizes. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.615162

Luijken, K., Lohmann, A., Alter, U., Gonzalez, J. C., Clouth, F. J., Fossum, J. L., Hesen, L., Huizing, A. H. J., Ketelaar, J., Montoya, A. K., Nab, L., Nijman, R. C. C., Vries, B. B. L. P. de, Tibbe, T. D., Wang, Y. A., & Groenwold, R. H. H. (2023). *Replicability of Simulation Studies for the Investigation of Statistical Methods: The RepliSims Project*. https://doi.org/10.48550/ARXIV.2307.02052

Martínez, S. F., & Huang, X. (2011). Epistemic groundings of abstraction and their cognitive dimension. *Philosophy of Science*, *78*(3), 490–511.

McCarley, J. S., Rose, L. E., Fischer, S., Haney, M. S., Hyk, A., Kuhn, K., Meredith, B., Moussaoui, J. R., Munoz Gomez Andrade, F., Pietrok, E. M., et al. (2023). Open science practices in the journal human factors: 2017–2022. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *67*, 1831–1836.

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do Frequency Representations Eliminate Conjunction Effects? An Exercise in Adversarial Collaboration. *Psychological Science*, *12*(4),

269–275. https://doi.org/10.1111/1467-9280.00350

Melloni, L., Mudrik, L., Pitts, M., Bendtz, K., Ferrante, O., Gorska, U., Hirschhorn, R., Khalaf, A., Kozma, C., Lepauvre, A., Liu, L., Mazumder, D., Richter, D., Zhou, H., Blumenfeld, H., Boly, M., Chalmers, D. J., Devore, S., Fallon, F., … Tononi, G. (2023). An adversarial collaboration protocol for testing contrasting predictions of global neuronal workspace and integrated information theory. *PloS One*, *18*(2), e0268577. https://doi.org/10.1371/journal.pone.0268577

Melloni, L., Mudrik, L., Pitts, M., & Koch, C. (2021). Making the hard problem of consciousness easier. *Science*, *372*(6545), 911–912. https://doi.org/10.1126/science.abj3259

Merkel, D. (2014). Docker: Lightweight linux containers for consistent development and deployment. *Linux Journal*, *2014*(239), 2.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., et al. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*(1), 719–748.

O'Kelly, M., Anisimov, V., Campbell, C., & Hamilton, S. (2017). Proposed best practice for projects that involve modelling and simulation: Proposed Best Practice for Modelling and Simulation. *Pharmaceutical Statistics*, *16*(2), 107–113. https://doi.org/10.1002/pst.1789

Pawel, S., Kook, L., & Reeve, K. (2023). Pitfalls and potentials in simulation studies: Questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method. *Biometrical Journal*, *66*(1). https://doi.org/10.1002/bimj.202200091

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo Experiments: Design and Implementation. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(2), 287–312. https://doi.org/10.1207/S15328007SEM0802_7

Popper, K. R. (1963). Science as falsification. *Conjectures and Refutations*, *1*(1963), 33–39.

Popper, K. R. (1988). *The open universe: An argument for indeterminism* (Vol. 2). Psychology Press.

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rakow, T. (2022). Adversarial Collaboration. In W. O'Donohue, A. Masuda, & S. Lilienfeld (Eds.), *Avoiding Questionable Research Practices in Applied Psychology* (pp. 359–377). Springer International Publishing. https://doi.org/10.1007/978-3-031-04968-2_16

Robitzsch, A. (2022). Comparing the Robustness of the Structural after Measurement (SAM)

Approach to Structural Equation Modeling (SEM) against Local Model Misspecifications with Alternative Estimation Approaches. *Stats*, *5*(3), 631–672. https://doi.org/10.3390/stats5030039

Rosseel, Y. (2012). Lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*, 1–36. https://doi.org/10.18637/jss.v048.i02

Rosseel, Y. (2020). Small sample solutions for structural equation modeling. In *Small sample size solutions* (pp. 226–238). Routledge.

Rosseel, Y., & Loh, W. W. (2022). A structural after measurement approach to structural equation modeling. *Psychological Methods*. https://doi.org/10.1037/met0000503

Savalei, V., & Kolenikov, S. (2008). Constrained versus unconstrained estimation in structural equation modeling. *Psychological Methods*, *13*(2), 150.

Slife, B. D., Wright, C. D., & Yanchar, S. C. (2016). Using operational definitions in research: A best-practices approach. *The Journal of Mind and Behavior*, 119–139.

Steenkamp, J.-B. E., & Maydeu-Olivares, A. (2023). Unrestricted factor analysis: A powerful alternative to confirmatory factor analysis. *Journal of the Academy of Marketing Science*, *51*(1), 86–113.

Supino, P. G. (2012). Overview of the research process. In P. G. Supino & J. S. Borer (Eds.), *Principles of research methodology: A guide for clinical investigators* (pp. 1–14). Springer New York. https://doi.org/10.1007/978-1-4614-3360-6_1

Thomopoulos, N. T. (2012). *Essentials of monte carlo simulation: Statistical methods for building simulation models*. Springer Science & Business Media.

Ulitzsch, E., Lüdtke, O., & Robitzsch, A. (2023). Alleviating estimation problems in small sample structural equation modeling—A comparison of constrained maximum likelihood, Bayesian estimation, and fixed reliability approaches. *Psychological Methods*, *28*(3), 527–557. https://doi.org/10.1037/met0000435

Ushey, K., & Wickham, H. (2024). *Renv: Project environments*. https://CRAN.R-project.org/package=renv

Van Driel, O. P. (1978). On various causes of improper solutions in maximum likelihood factor analysis. *Psychometrika*, *43*, 225–243.

Vaughan, D., & Dancho, M. (2022). *Furrr: Apply mapping functions in parallel using futures*. https://CRAN.R-project.org/package=furrr

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., & Henry, L. (2023). *Purrr: Functional programming tools*.

https://purrr.tidyverse.org/

Wolins, L. (1995). A monte carlo study of constrained factor analysis using maximum likelihood and unweighted least squares. *Educational and Psychological Measurement*, *55*(4), 545–557.

Xie, Y. (2024). *Knitr: A general-purpose package for dynamic report generation in r*. https://yihui.org/knitr/

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1.

Zhu, H. (2024). *kableExtra: Construct complex table with 'kable' and pipe syntax*. http://haozhu233.github.io/kableExtra/

Zhu, H., Travison, T., Tsai, T., Beasley, W., Xie, Y., Yu, G., Laurent, S., Shepherd, R., Sidi, Y., Salzer, B., Gui, G., Fan, Y., Murdoch, D., Arel-Bundock, V., & Evans, B. (2024). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. https://cran.r-project.org/web/packages/kableExtra/index.html

Appendix A: Simulation Protocol

Here, the full simulation protocol of my simulation studies conducted individually prior to collaboration as well as the follow up study I conducted in light of the collaboration with Kosanke after the first round of conducting and evaluating our individual studies is presented. It is based on the preregistration of my individual studies (Kriegmair, 2024) and outlines all deviations from it.

Preregistration template designed by: Björn S. Siepe, František Bartoš, Tim P. Morris, Anne-Laure Boulesteix, Daniel W. Heck, and Samuel Pawel

1. General Information

1.1 What is the title of the project?

Comparing a Structural After Measurement (SAM) Approach to Standard Structural Equation Model (SEM) Estimation

1.2 Who are the current and future project contributors?

Valentin Kriegmair

1.3 Provide a description of the project.

The studies registered were part of an adversarial collaboration project. The aim was to conceptually (only in part) replicate the results obtained by Dhaene & Rosseel (2023) and Rosseel & Loh (2022). I set out to evaluate the performance of a Structural After Measurement (SAM) approach for estimating structural equation models (SEM) in comparison to standard SEM estimation methods. This served as the basis for the adversarial collaboration with another researcher who evaluated the same research question from the perspective of a conceptual replication of the (in part contradicting) results obtained by Robitzsch (2022). However, the following only describes the first (conceptual) replication.

1.4 Did any of the contributors already conduct related simulation studies on this specific question?

No prior related simulation studies have been conducted by the contributors.

2. Aims

Structural After Measurement (SAM) is an estimation method for structural equation models that consists of a stepwise estimation of the measurement and structural parts of a model. The research questions of the current simulation were:

1. How do SAM and traditional SEM methods (including ML and ULS) compare in terms of bias, Mean Squared Error (MSE), and convergence rates in small to moderate samples?
2. What is the impact of model misspecifications, such as residual correlations and cross-loadings, on the performance of SAM compared to traditional SEM methods?

3. Data-Generating Mechanism

3.1 Study 1

In study 1 (conceptually replicating Rosseel & Loh (2022)) data was generated parametrically. Four different population structural equation models (SEM) with five latent variables and three continuous indicators per facotr based on the following matrices were simulated:

- B as $M \times M$ matrix representing latent regression coefficients with all $b = 0.1$.
  - Model 1.1 and 1.2:

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0.1 & 0.1 & 0 & 0.1 & 0 \\ 0.1 & 0.1 & 0 & 0 & 0 \\ 0.1 & 0 & 0.1 & 0.1 & 0 \end{bmatrix}$$

  - Model 1.3 in deviation from the preregistration with a reversed effect between latent factors f3 and f4 to introduce another realistic and more severe misspecification to show the potential of SAM in most challenging conditions:

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0.1 & 0.1 & 0 & 0 & 0 \\ 0.1 & 0.1 & 0.1 & 0 & 0 \\ 0.1 & 0 & 0.1 & 0.1 & 0 \end{bmatrix}$$

  - Model 1.4 in deviation from the preregistration with a bidirectional structural relation

between f3 and f4 specified as only one directional instead of just reversing the effect to investigate a different type of misspecification:

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0.1 & 0.1 & 0 & 0.1 & 0 \\ 0.1 & 0.1 & 0.1 & 0 & 0 \\ 0.1 & 0 & 0.1 & 0.1 & 0 \end{bmatrix}$$

- Ψ as $M \times M$ as diagonal matrix representing the residual variances in deviation from the preregistration not adjusted for the varying structural relations. This was only updated in the joint study (study 3) to adjust residual variances of all endogenous factors to accurately reflect the number of regressors
  - Model 1.1, 1.2, 1.3, and 1.4:

$$\Psi = \begin{bmatrix} 1.0 & 0 & 0 & 0 & 0 \\ 0 & 1.0 & 0 & 0 & 0 \\ 0 & 0 & 1.0 & 0 & 0 \\ 0 & 0 & 0 & 1.0 & 0 \\ 0 & 0 & 0 & 0 & 1.0 \end{bmatrix}$$

- Λ as $P \times M$ matrix representing factor loadings.

- Model 1.1, 1.3 and 1.4:

$$\Lambda = \begin{bmatrix} 1.0 & 0 & 0 & 0 & 0 \\ 0.7 & 0 & 0 & 0 & 0 \\ 0.7 & 0 & 0 & 0 & 0 \\ 0 & 1.0 & 0 & 0 & 0 \\ 0 & 0.7 & 0 & 0 & 0 \\ 0 & 0.7 & 0 & 0 & 0 \\ 0 & 0 & 1.0 & 0 & 0 \\ 0 & 0 & 0.7 & 0 & 0 \\ 0 & 0 & 0.7 & 0 & 0 \\ 0 & 0 & 0 & 1.0 & 0 \\ 0 & 0 & 0 & 0.7 & 0 \\ 0 & 0 & 0 & 0.7 & 0 \\ 0 & 0 & 0 & 0 & 1.0 \\ 0 & 0 & 0 & 0 & 0.7 \\ 0 & 0 & 0 & 0 & 0.7 \end{bmatrix}$$

- Model 1.2: cross loadings will be set to be 10% lower than the factor loadings: $\Lambda_{ik,jk} = 0.63 = 0.9 \times 0.7$. They will be generated by the following elements in Λ: (2, 2), (5, 3), (8, 4), (11, 5), (14, 1).

- Θ as a $P \times P$ matrix representing the residual variances and covariances of the indicators.
  - Model 1.1, 1.2 and 1.4: The diagonal generated as:

$$\Theta^* = \text{Var}(\eta)\Lambda^T \times \frac{1}{r-1}$$

    (where r is the reliability of the indicators) and 0 on all off-diagonal elements
  - Model 1.3:
    - Θ^* on the diagonal.
    - Correlated residuals generated between specific indicator pairs: for $i = (2, 5, 8, 11, 14)$ and $i' = (3, 6, 9, 12, 15)$, and for each $k = 1, \ldots, 4$ and $l = k + 1, \ldots, 5$, the entries (i_k, i'_l) and (i'_l, i_k) in Θ are set to $0.6 \times \min \Theta^*$, ensuring correlated errors among selected indicator pairs without exceeding a 0.6 correlation coefficient.

3.1.2 Study 2

For study 2, again, different five-factor population models with three continious indicators per facotr were generated parametrically . Further, the different models of study 2 were used for different simulation settings resulting in two sub-studies 2.1 and 2.2 (see simulation settings).

- B as $M \times M$ matrix representing latent regression coefficients with varying parameter size defined by two conditions of endogenous factor variance explained by the exogenous factors (low: $R^2 = 0.1$ or medium: $R^2 = 0.4$ see below under factor):
  - Model 2.1 and 2.2:

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \beta_{\eta_4,\eta_1} & \beta_{\eta_4,\eta_2} & 0 & 0 & 0 \\ 0 & \beta_{\eta_5,\eta_2} & \beta_{\eta_5,\eta_3} & \beta_{\eta_5,\eta_4} & 0 \end{bmatrix}$$

- Λ as $P \times M$ matrix representing factor loadings of indicators on the latent factors.
  - Model 2.1:

$$\Lambda = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0.9 & 0 & 0 & 0 & 0 \\ 0.8 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0.9 & 0 & 0 & 0 \\ 0 & 0.8 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0.9 & 0 & 0 \\ 0 & 0 & 0.8 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0.9 & 0 \\ 0 & 0 & 0 & 0.8 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0.9 \\ 0 & 0 & 0 & 0 & 0.8 \end{bmatrix}$$

  - Model 2.2 with cross-loadings either in the exogenous ($\lambda_{6,3}$), endogenous ($\lambda_{12,5}$) or both parts of the model. Which cross loading was present depended on the misspecification simulation factor. The specific magnitude of the endogenous ($\lambda_{12,5}$)

loading depended on R^2 (see under 3.2.2):

$$\Lambda = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0.9 & 0 & 0 & 0 & 0 \\ 0.8 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0.9 & 0 & 0 & 0 \\ 0 & 0.8 & \lambda_{6,3} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0.9 & 0 & 0 \\ 0 & 0 & 0.8 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0.9 & 0 \\ 0 & 0 & 0 & 0.8 & \lambda_{12,5} \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0.9 \\ 0 & 0 & 0 & 0 & 0.8 \end{bmatrix}$$

- Θ as a $P \times P$ matrix representing the residual variances and covariances of the indicators. This was computed as the portion of the indicator's total variance that is not explained by the latent factors, after accounting for the strength and reliability of its relationship to these factors (factor loadings), as well as the effects of regressions between the latent factors themselves.
  - Model 2.1: The diagonal of Θ generated as:

$$\Theta^* = \text{Var}(\eta)\Lambda^T \times \frac{1}{r-1}$$

    (where r is the reliability of the indicators) and 0 on all off-diagonal elements
  - Model 2.2:
    * Θ^* on the diagonal.
    * Correlated residuals generated between specific indicator pairs in either the endogenous, exogenous or both parts of the model.

    Thus depending on the simulation setting either:

    * $\Theta_{8,9}$, $\Theta_{9,8}$ (exogenous part)
    * $\Theta_{14,15}$ and $\Theta_{15,14}$ (endogenous part)
    * $\Theta_{8,9}$, $\Theta_{9,8}$, $\Theta_{14,15}$ and $\Theta_{15,14}$ (both parts)

were set $0.6 \times \min \Theta^*$, ensuring correlated errors among selected indicator pairs without exceeding a 0.6 correlation coefficient:

3.1.3 Study 3

For study 3, again, four different five-factor population models with three indicators per factor were generated parametrically with B as $M \times M$ matrix representing latent regression coefficients with all $b = 0.1$ for all models in study 3:

$$\Psi = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0.1 & 0.1 & 0 & 0.1 & 0 \\ 0.1 & 0.1 & 0 & 0 & 0 \\ 0.1 & 0 & 0.1 & 0.1 & 0 \end{bmatrix}$$

and Ψ as $M \times M$ as diagonal matrix (0 on the off diagonal) representing variances of the factors with $1 - kb^2$ on the diagonal where k is the number of latent regressor per factor and b the regression coefficients (0.1) for all models in study 3. Each model in study 3 included either cross loadings or correlated residual errors in the measurement model based on Λ and Θ (constructed as in study 1) but these modifications in the measurement models could be either positive or negative.

3.2 Factors of the Data-Generating Mechanism

3.2.1 Study 1 The first study modulated the following factors:

- Different misspecifications of the population model where the population model varies between the different models (1.1, 1.2, 1.3, 1.4) as described above, while the analysis model remains specified as model 1.1.
- Sample sizes of small (N = 100), medium (N = 400), or large (N = 6400)
- Indicator reliability of low (.3), moderate (.5), or high (.7)

3.2.2 Study 2 The second study modulated the following factors of the data generating process across both studies:

- Sample sizes of small (N = 100), medium (N = 400), or large (N = 6400)
- Variance explained (R^2) of the endogenous factor variance explained by the exogenous factors: low ($R^2 = 0.1$) or medium ($R^2 = 0.4$)
- Indicator reliability of three indicators per factor: *all high* (.8), *all low* (.5), *average low* (.5) varying between .7 to .3 with the highest reliability for the scaling indicator.

- Sample sizes of small (N = 100), medium (N = 400), or large (N = 6400)
- Deviating from the preregistration distribution (normal vs. non-normal) was not considered in the simulation settings to limit the scope of the study.
- Measurement misspecifications of a residual covariance and a factor loading either in the exogenous, endogoneous or both parts of the model (in deviation from the preregistration without additional structural misspecifications and only three modulations to limit the scope of the study):
- Number of measurement blocks (how many separate measurement models are fitted in the first step of SAM) of either a separate measurement model per latent variable (b = k = 5) or one joint measurement model for all exogenous variables (b = 3)

In deviation from the preregistration, additionally all models in study 2 were estimated including stuructral specifications that were not present in the population model to investigate the performance of the methods on recovering falsely specified absent structural relations.

3.2.3 Study 3 The third study modulated the following factors of the data generating process:

- Sample sizes of N = 50, N = 100, N = 250 and N = 400.
- Indicator reliability of low (.3), moderate (.5), or high (.7)

3.3 Simulation Conditions

- Study 1: in deviation from the preregistration only one estimattion model was considered to limit the scope of the study resulting in 36 conditions (4 population models x 3 sample sizes x 3 reliabilities)
- Study 2.1 (4 population models x 3 sample sizes x 2 R^2 x 3 reliabilities x 2 measurement blocks = 144 conditions) (in dviation from the preregistration the misspecifications were reduced and counted here as differnet population odels as well)

4. Estimands and Targets

Estimated structural model parameters (path coefficients) represented the estimands of interest.

5. Methods

Both studies will compare four different estimation methods for SEMs:

- Traditional SEM: (structural and measurement model estimated simultaneously) (rationale: the current standard approach in SEM estimation serving as a baseline with maximum likelihood (ML)):
- SAM: (separating the estimation of the measurement and structural model to alleviate the potential for propagation of bias from (e.g. misspecified) measurement part to the structural part of the model)
  - Local SAM (Uses summary statistics from the measurement model to derive the model-implied mean vector and variance-covariance matrix of latent variables. These statistics are then utilized to estimate the structural parameters. A mapping matrix (M) is used to transform the observed data into the latent variable space. It can be estimated using different methods.)
    - With ML mapping matrix (Akin to a factor score approach Bartlett (1938))
    - With ULS mapping matrix (Uses the Moore-Penrose pseudoinverse, suitable for scenarios with complex or underdetermined systems, where the K matrix is rank-deficient but requires adjustments for structural constraints.)
  - Global SAM (rationale: Fixing the parameters obtained from the measurement model in the first step, and then using them as constants in the full SEM during the second step. Suitable for models where local SAM is impractical due to higher-order latent variables or rank deficiencies in λ.)

Traditional SEM as well as both steps in the SAM approach will be estimated using Maximum Likelihood (ML) using `lavaan` (Rosseel, 2012) in R 4.4 (R Core Team, 2023).

6. Performance Measures

Across both studies the following performance measures were captured:

- Convergence rates: Proportions of observed data sets that successfully converged for each estimation method detected using `lavaan`.
- In deviation from the preregistration also impropper solutions of converged models showing negative variances as the only type of improper solution present were computed.
- Rrelative biases: Average difference between an estimate and its true value, normalized by the true value, assessed across all path coefficients: $\frac{\bar{T}-\theta}{\theta}$
- Absolute biases: (in deviation from the preregistration this measure as it might be more intuitive and applicable for study 1 and 3 with invariant regression weights): $(\bar{T}-\theta)$
- Root Mean Squared Errors (RMSE): Calculated as the square root of the average squared difference between an estimate and its true value, evaluated under conditions of model

misspecification: $(\sqrt{\frac{1}{K}\sum_{k=1}^{K}(T_k-\theta)^2})$ where T_k is the estimated parameter, $\bar{T}$ the mean of the estimated parameters and θ the true parameter value, and K is the number of replications computed.

- Relative Root Mean Squared Errors (RRMSE) in deviation from preregistration for better comparability in study 2 under varying regression weights:$\sqrt{\frac{(\bar{T}-\theta)^2+S_T^2}{\theta^2}}$
- Empirical coverage levels of 95% confidence intervals (CIs): Proportion of observed data sets where the constructed CIs included the true value. (Not reported to limit the scope)

7. Monte Carlo Uncertainty of the Estimated Performance Measures

Monte Carlo uncertainty was calculated (manually in deviation from the preregistration) for the absolute and relative metrics: $\sqrt{\frac{S_T^2}{K}}$ and $\sqrt{\frac{S_T^2}{K\theta^2}}$ for bias and relative bias, and $\sqrt{\frac{K-1}{K}\sum_{j=1}^{K}\left(\text{RMSE}_{(j)}-\text{RMSE}\right)^2}$ and $\sqrt{\frac{K-1}{K}\sum_{j=1}^{K}\left(rRMSE_{(j)}-rRMSE\right)^2}$ for RMSE and relative RMSE.

8. Simulation Repetitions

- Replicating Rosseel & Loh (2022) study 1 consisted of 5000 repetitions per condition.
- Replicating Dhaene & Rosseel (2023) study 2 will consisted of 10000 repetitions per condition.
- Study 3 entailed 5000 repitiions as this resulted in sufficiently small Monte Carlo standard errors for the performance measures.

9. Missing Values due to non-convergence or other reasons

As mentioned above convergence rates were captured and only converged propper solutions were used for performance measure computation.

10. Software and Libraries

The simulation was set up and conducted in R Core Team (2023) using `lavaan` (Rosseel, 2012) for generating data and estimation. The `furrr` (Vaughan & Dancho, 2022) package for parallel simulation execution. A full list of libraries and dependencies can be found on GitHub

11. Computational Environment

The simulations were conducted using the TARDIS high-performance computing cluster at the Max Planck Institute for Human Development. The computational environment was set up in R, utilizing a suite of packages for analysis and parallel computing. Key libraries included:

- Analysis and Data Manipulation Packages: `MASS`, `dplyr`, `tidyr`, `lavaan`, `purrr`, and `Matrix`.
- Parallel Computing Packages: `future`, `furrr`, `parallel`, `future` and `batchtools`.

12. Reproducibility

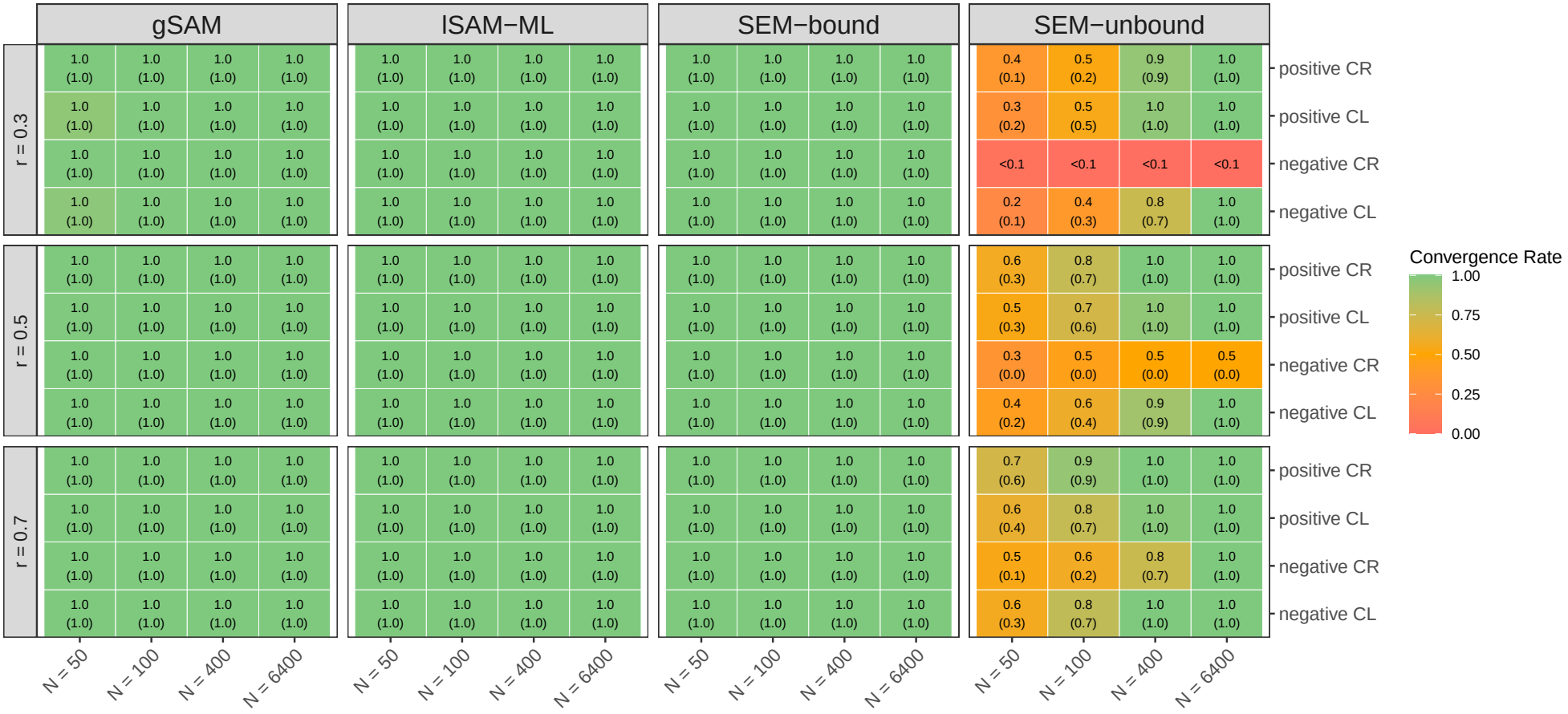
The code of the simulation was made available on [GitHub] (https://github.com/valentinkm/AdversarialSimulation). A pre-generated list of seeds was used for all replications to ensure reproducibility and avoid synchronization in parallelized computations. As a examplary replication the simulation can be reproduced in this GitHub action here.

Appendix B: Supplementary Figures

In the following, additional figures are presented that were not included in the main text and provide further insights into the simulation results.

Figure B1

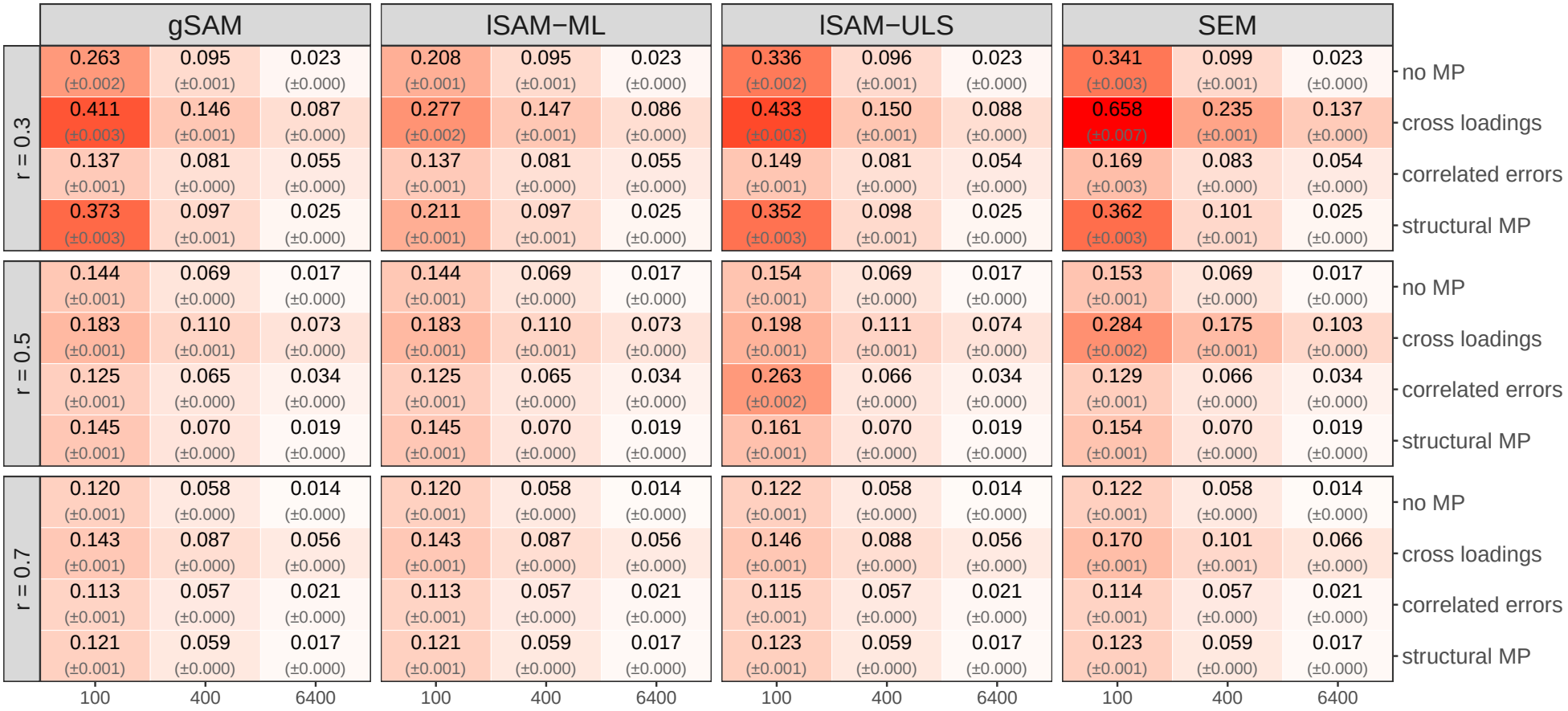
Convergence Rate and Rate of Proper Solutions in Study 3



Note. Convergence and proper solutions (in parentheses) rates across sample sizes (N), reliability (r), and model misspecifications for global SAM (gSAM), local SAM with Maximum Likelihood (lSAM-ML), Unweighted Least Squares (lSAM-ULS), and SEM. Graphic adapted from Dhaene & Rosseel (2023).

Figure B2

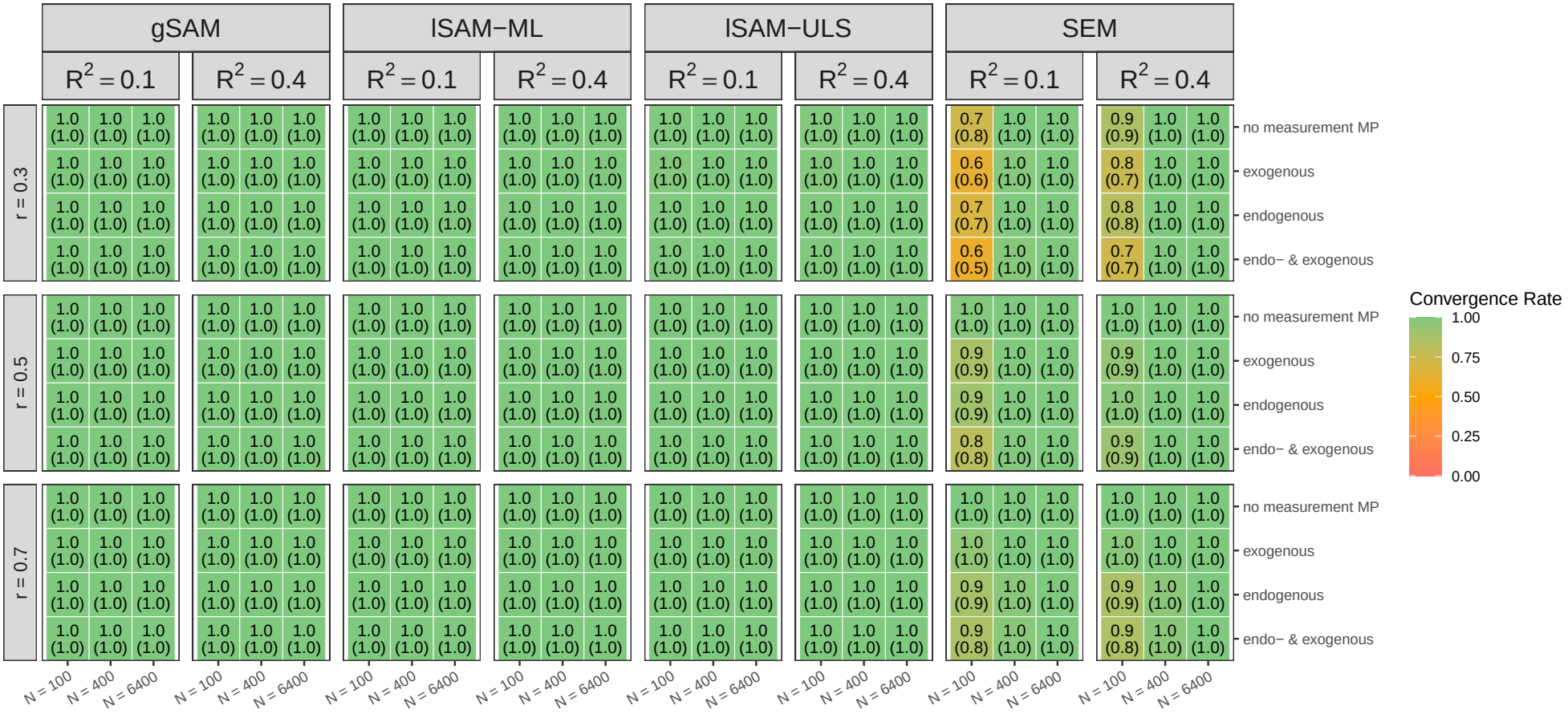
Mean Average Root Mean Squared Error (RMSE) of Regression Parameters in Study 1



Note. Mean RMSE averaged (in absolute values) over all parameters in one model for sample sizes (N), reliability (r), and misspecifications for global SAM (gSAM), local SAM with Maximum Likelihood (lSAM-ML), Unweighted Least Squares (lSAM-ULS) and SEM. Monte Carlo Standard Error (MCSE) in parentheses. Graphic adapted from Dhaene & Rosseel (2023).

Figure B3

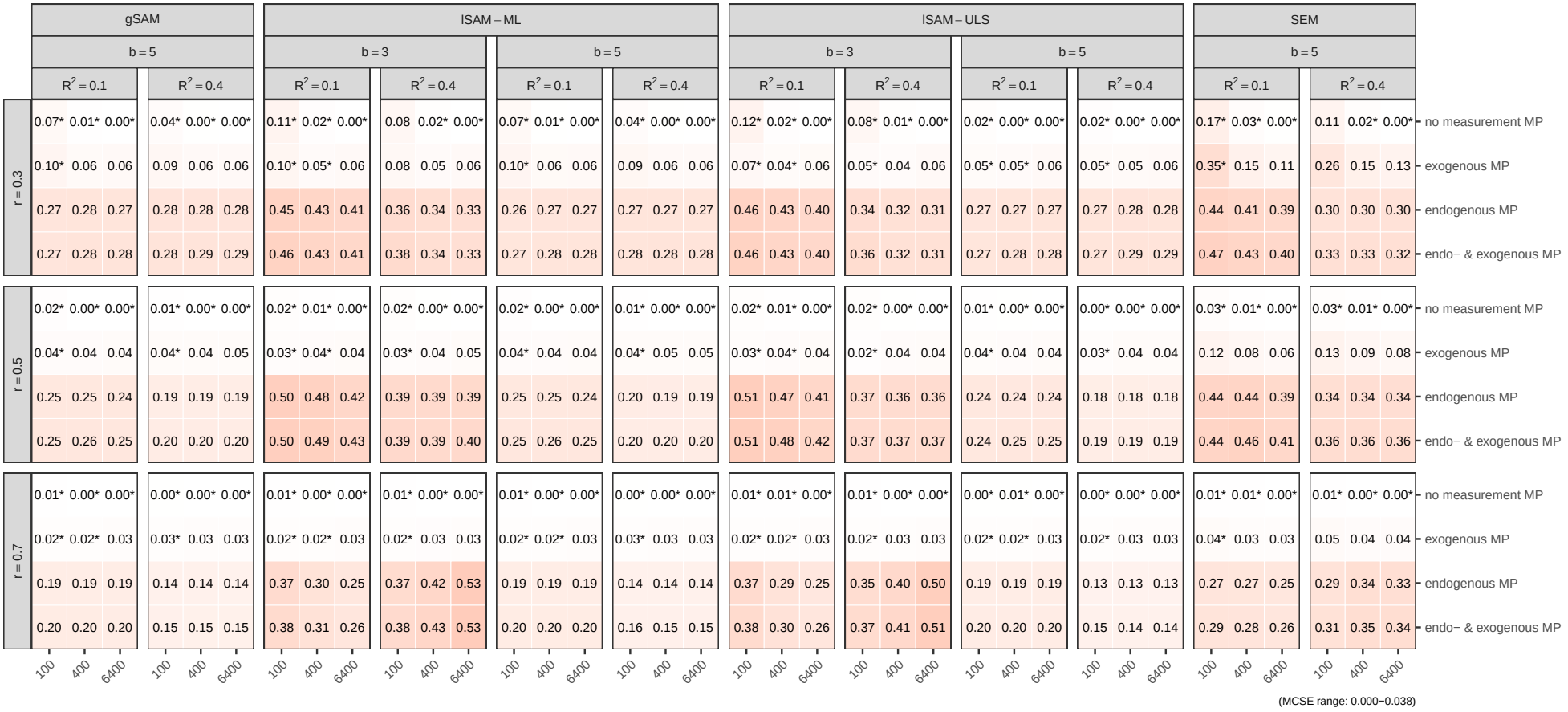
Convergence Rate and Rate of Proper Solutions in Study 2



Note. Convergence and proper solutions (in parentheses) rates across sample sizes (N), reliability (r), and model misspecification location for global SAM (gSAM), local SAM with Maximum Likelihood (lSAM-ML), Unweighted Least Squares (lSAM-ULS), and SEM. Graphic adapted from Dhaene & Rosseel (2023).

Figure B4

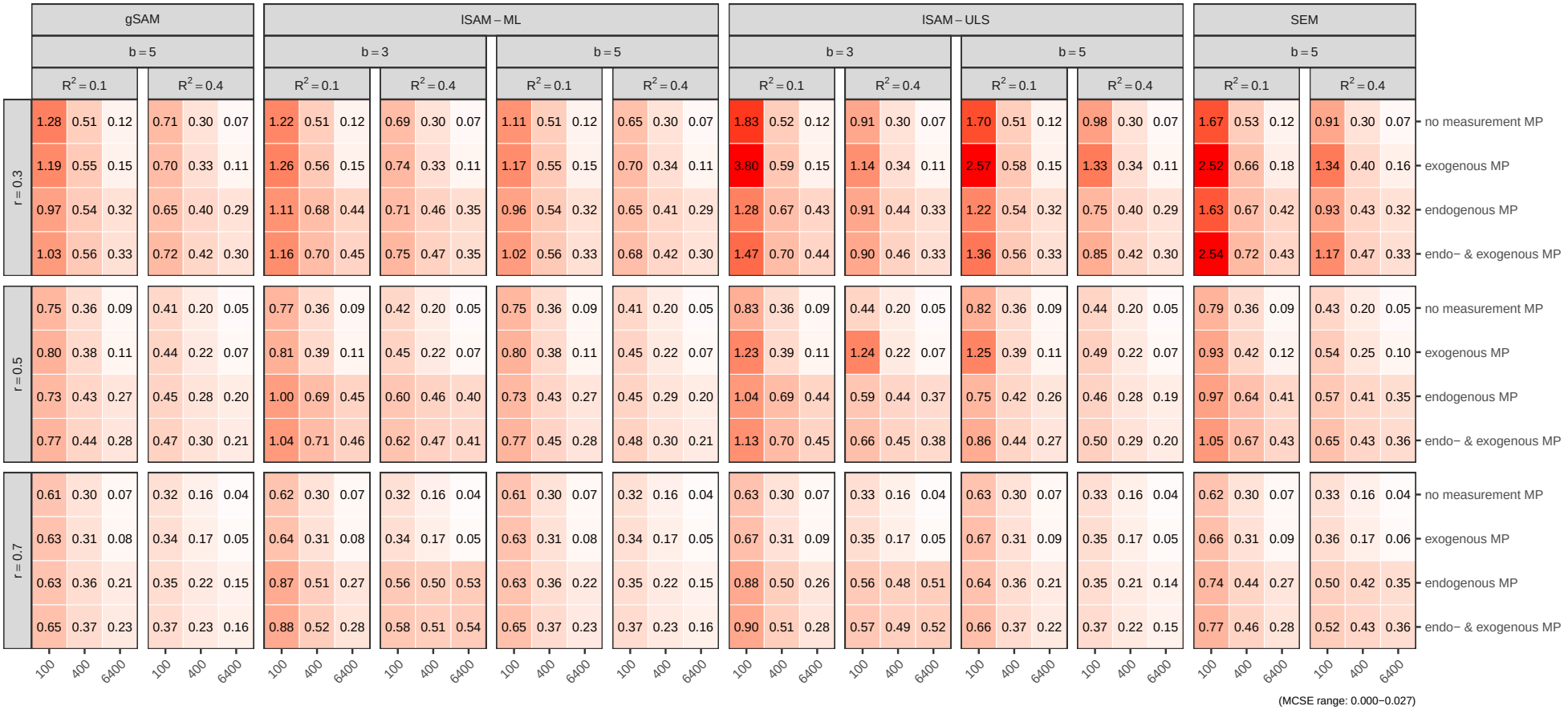
Relative Bias of Regression Parameters in Study 2



Note. Mean relative bias averaged (in absolute values) over all parameters in one model for sample sizes (N), reliability (r), and misspecifications for global SAM (gSAM), local SAM with Maximum Likelihood (lSAM-ML), Unweighted Least Squares (lSAM-ULS) and SEM. * indicating Monte-Carlos Standard Error (MCSE) above 10% of the estimate. Graphic adapted from Dhaene & Rosseel (2023).

Figure B5

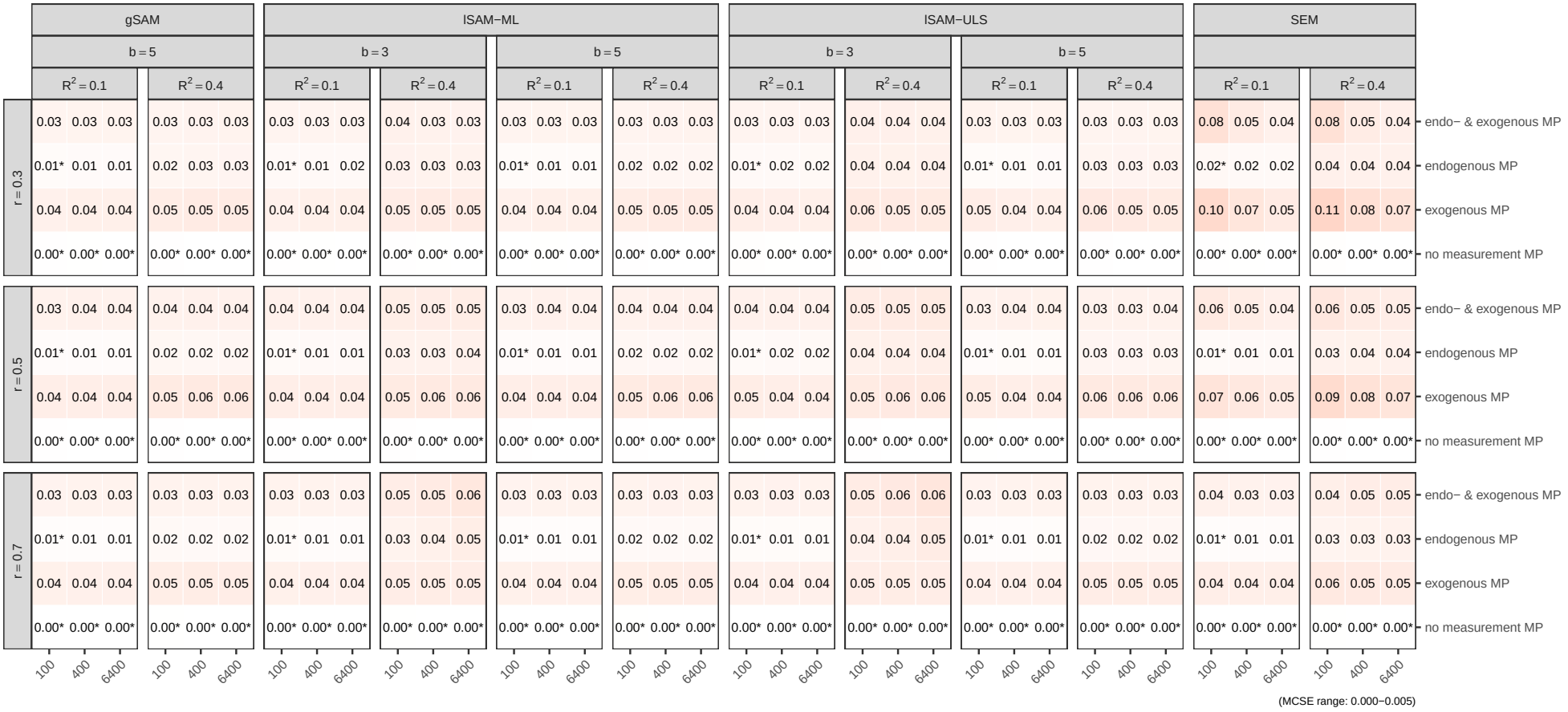
Relative RMSE of Regression Parameters in Study 2



Note. Mean relative RMSE averaged (in absolute values) over all parameters in one model for sample sizes (N), reliability (r), and misspecifications for global SAM (gSAM), local SAM with Maximum Likelihood (lSAM-ML), Unweighted Least Squares (lSAM-ULS) and SEM. * indicating Monte-Carlos Standard Error (MCSE) above 10% of the estimate. Graphic adapted from Dhaene & Rosseel (2023).

Figure B6

Bias of Misspecified Regression Parameters in Study 2



Note. Mean bias of parameters absent in the population and misspecified in the analysis model for sample sizes (N), reliability (r), and misspecifications for global SAM (gSAM), local SAM with Maximum Likelihood (lSAM-ML), Unweighted Least Squares (lSAM-ULS) and SEM. * indicating Monte-Carlos Standard Error (MCSE) above 10% of the estimate. Graphic adapted from Dhaene & Rosseel (2023).

Appendix C: Detailed Error and Warning Messages

Here, all warning and error messages raised during the studies are listed (see Table C1) and shown how often they occurred under various fitting conditions (see Table C2).

Table C1

List of Unique Warnings and Errors

ID	Message
1	lavaan WARNING: some estimated ov variances are negative
2	lavaan WARNING: the optimizer warns that a solution has NOT been found!
3	lavaan WARNING: the optimizer (NLMINB) claimed the model converged, but not all elements of the gradient are (near) zero; the optimizer may not have found a local solution use check.gradient = FALSE to skip this check.
4	lavaan WARNING: some estimated lv variances are negative
5	lavaan WARNING: some estimated ov variances are negative, lavaan WARNING: some estimated lv variances are negative
6	number of items to replace is not a multiple of replacement length
7	lavaan WARNING: Could not compute standard errors! The information matrix could not be inverted. This may be a symptom that the model is not identified., lavaan WARNING: some estimated ov variances are negative
8	lavaan WARNING: covariance matrix of latent variables is not positive definite; use lavInspect(fit, ”cov.lv”) to investigate.
9	lavaan WARNING: The variance-covariance matrix of the estimated parameters (vcov) does not appear to be positive definite! The smallest eigenvalue is smaller than or close to zero. This may be a symptom that the model is not identified., lavaan WARNING: some estimated ov variances are negative
10	lavaan WARNING: Could not compute standard errors! The information matrix could not be inverted. This may be a symptom that the model is not identified., lavaan WARNING: some estimated lv variances are negative
11	lavaan WARNING: Could not compute standard errors! The information matrix could not be inverted. This may be a symptom that the model is not identified., lavaan WARNING: some estimated ov variances are negative, lavaan WARNING: some estimated lv variances are negative
12	lavaan WARNING: some estimated ov variances are negative, lavaan WARNING: covariance matrix of latent variables is not positive definite; use lavInspect(fit, ”cov.lv”) to investigate.

Table C1

List of Unique Warnings and Errors (continued)

ID	Message
13	lavaan WARNING: Could not compute standard errors! The information matrix could not be inverted. This may be a symptom that the model is not identified.
14	lavaan WARNING: The variance-covariance matrix of the estimated parameters (vcov) does not appear to be positive definite! The smallest eigenvalue is smaller than or close to zero. This may be a symptom that the model is not identified., lavaan WARNING: covariance matrix of latent variables is not positive definite; use lavInspect(fit, "cov.lv") to investigate.
15	lavaan WARNING: Could not compute standard errors! The information matrix could not be inverted. This may be a symptom that the model is not identified., lavaan WARNING: covariance matrix of latent variables is not positive definite; use lavInspect(fit, "cov.lv") to investigate.
16	lavaan WARNING: Could not compute standard errors! The information matrix could not be inverted. This may be a symptom that the model is not identified., lavaan WARNING: some estimated ov variances are negative, lavaan WARNING: covariance matrix of latent variables is not positive definite; use lavInspect(fit, "cov.lv") to investigate.
17	lavaan WARNING: The variance-covariance matrix of the estimated parameters (vcov) does not appear to be positive definite! The smallest eigenvalue is smaller than or close to zero. This may be a symptom that the model is not identified.
18	lavaan WARNING: The variance-covariance matrix of the estimated parameters (vcov) does not appear to be positive definite! The smallest eigenvalue is smaller than or close to zero. This may be a symptom that the model is not identified., lavaan WARNING: some estimated ov variances are negative, lavaan WARNING: some estimated lv variances are negative

Note. This table lists all unique warnings and errors encountered during the simulation studies.

Table C2

Summary of Warnings and Errors by Condition with ID for All Studies

Study	Misspecification	N	r	Method	Type	Count	ID
Study 1	correlated errors	100	0.3	SEM	Warning	6860	1
Study 1	cross loadings	100	0.5	SEM	Warning	3575	1
Study 1	correlated errors	100	0.5	SEM	Warning	2923	1
Study 1	cross loadings	100	0.7	SEM	Warning	2903	1

Table C2

Summary of Warnings and Errors by Condition with ID for All Studies (continued)

Study	Misspecification	N	r	Method	Type	Count	ID
Study 1	cross loadings	100	0.3	SEM	Warning	2769	1
Study 1	no measurement MP	100	0.3	SEM	Warning	2700	1
Study 1	structural MP	100	0.3	SEM	Warning	2577	1
Study 1	cross loadings	100	0.3	SEM	Warning	2037	2
Study 1	no measurement MP	100	0.3	SEM	Warning	1258	2
Study 1	structural MP	100	0.3	SEM	Warning	1133	2
Study 1	cross loadings	100	0.3	SEM	Warning	729	3
Study 1	correlated errors	100	0.3	SEM	Warning	692	3
Study 1	correlated errors	100	0.7	SEM	Warning	688	1
Study 1	correlated errors	400	0.3	SEM	Warning	606	1
Study 1	correlated errors	100	0.3	SEM	Warning	507	2
Study 1	cross loadings	400	0.3	SEM	Warning	450	1
Study 1	cross loadings	400	0.5	SEM	Warning	429	1
Study 1	cross loadings	100	0.3	SEM	Warning	417	4
Study 1	cross loadings	400	0.7	SEM	Warning	248	1
Study 1	no measurement MP	100	0.3	SEM	Warning	242	3
Study 1	structural MP	100	0.3	SEM	Warning	223	3
Study 1	cross loadings	100	0.5	SEM	Warning	203	2
Study 1	no measurement MP	100	0.5	SEM	Warning	197	1
Study 1	structural MP	100	0.5	SEM	Warning	183	1
Study 1	cross loadings	100	0.3	lSAM-ULS	Warning	150	1
Study 1	cross loadings	100	0.3	SEM	Warning	146	5
Study 1	structural MP	100	0.3	lSAM-ULS	Warning	62	1
Study 1	no measurement MP	100	0.3	lSAM-ULS	Warning	52	1
Study 1	cross loadings	100	0.3	gSAM	Warning	50	4
Study 1	structural MP	100	0.3	SEM	Warning	50	4
Study 1	cross loadings	100	0.5	SEM	Warning	42	3
Study 1	cross loadings	100	0.3	lSAM-ULS	Error	38	6
Study 1	no measurement MP	100	0.3	SEM	Warning	29	4
Study 1	no measurement MP	100	0.3	lSAM-ULS	Error	25	6
Study 1	structural MP	100	0.3	lSAM-ULS	Error	24	6
Study 1	cross loadings	100	0.3	SEM	Warning	23	7
Study 1	cross loadings	400	0.3	SEM	Warning	15	2
Study 1	no measurement MP	100	0.3	SEM	Warning	14	7

Table C2

Summary of Warnings and Errors by Condition with ID for All Studies (continued)

Study	Misspecification	N	r	Method	Type	Count	ID
Study 1	cross loadings	100	0.3	gSAM	Error	14	6
Study 1	no measurement MP	100	0.3	SEM	Warning	12	5
Study 1	structural MP	100	0.3	SEM	Warning	11	5
Study 1	structural MP	100	0.3	SEM	Warning	9	7
Study 1	cross loadings	100	0.7	SEM	Warning	7	2
Study 1	structural MP	100	0.7	SEM	Warning	7	1
Study 1	cross loadings	100	0.3	SEM	Warning	5	8
Study 1	no measurement MP	100	0.5	SEM	Warning	4	2
Study 1	correlated errors	100	0.5	SEM	Warning	4	3
Study 1	no measurement MP	100	0.3	SEM	Warning	3	9
Study 1	no measurement MP	100	0.7	SEM	Warning	3	1
Study 1	cross loadings	100	0.3	SEM	Warning	3	10
Study 1	cross loadings	100	0.3	SEM	Warning	3	11
Study 1	no measurement MP	100	0.3	gSAM	Error	2	6
Study 1	no measurement MP	100	0.3	gSAM	Warning	2	4
Study 1	no measurement MP	400	0.3	SEM	Warning	2	1
Study 1	cross loadings	100	0.3	SEM	Warning	2	12
Study 1	cross loadings	100	0.3	lSAM-ULS	Warning	2	3
Study 1	cross loadings	100	0.5	SEM	Warning	2	4
Study 1	cross loadings	400	0.3	SEM	Warning	2	3
Study 1	correlated errors	100	0.5	lSAM-ULS	Error	2	6
Study 1	correlated errors	100	0.5	lSAM-ULS	Warning	2	1
Study 1	structural MP	100	0.3	SEM	Warning	2	9
Study 1	structural MP	100	0.3	gSAM	Error	2	6
Study 1	structural MP	100	0.3	gSAM	Warning	2	4
Study 1	structural MP	100	0.5	lSAM-ULS	Error	2	6
Study 1	no measurement MP	100	0.3	SEM	Warning	1	11
Study 1	no measurement MP	100	0.3	lSAM-ULS	Warning	1	3
Study 1	no measurement MP	100	0.5	lSAM-ULS	Warning	1	1
Study 1	cross loadings	100	0.3	SEM	Warning	1	9
Study 1	cross loadings	100	0.3	gSAM	Warning	1	8
Study 1	cross loadings	100	0.5	lSAM-ULS	Error	1	6
Study 1	cross loadings	100	0.5	lSAM-ULS	Warning	1	1
Study 1	correlated errors	100	0.3	SEM	Warning	1	7

Table C2

Summary of Warnings and Errors by Condition with ID for All Studies (continued)

Study	Misspecification	N	r	Method	Type	Count	ID
Study 1	correlated errors	100	0.3	lSAM-ULS	Warning	1	1
Study 1	correlated errors	100	0.5	SEM	Warning	1	2
Study 1	correlated errors	400	0.5	SEM	Warning	1	1
Study 1	structural MP	100	0.3	SEM	Warning	1	10
Study 1	structural MP	100	0.3	lSAM-ULS	Warning	1	3
Study 1	structural MP	100	0.3	lSAM-ULS	Warning	1	2
Study 1	structural MP	100	0.5	lSAM-ULS	Warning	1	1
Study 2	endo- & exogenous MP	100	0.3	SEM	Warning	5265	1
Study 2	exogenous MP	100	0.3	SEM	Warning	4622	1
Study 2	endogenous MP	100	0.3	SEM	Warning	3615	1
Study 2	endo- & exogenous MP	100	0.7	SEM	Warning	2904	1
Study 2	endo- & exogenous MP	100	0.5	SEM	Warning	2743	1
Study 2	no measurement MP	100	0.3	SEM	Warning	2701	1
Study 2	endogenous MP	100	0.7	SEM	Warning	2336	1
Study 2	exogenous MP	100	0.5	SEM	Warning	1814	1
Study 2	endo- & exogenous MP	100	0.3	SEM	Warning	1625	2
Study 2	exogenous MP	100	0.3	SEM	Warning	1624	2
Study 2	endogenous MP	100	0.5	SEM	Warning	1252	1
Study 2	endogenous MP	100	0.3	SEM	Warning	1211	2
Study 2	no measurement MP	100	0.3	SEM	Warning	1121	2
Study 2	exogenous MP	100	0.7	SEM	Warning	702	1
Study 2	endo- & exogenous MP	100	0.3	SEM	Warning	675	3
Study 2	endo- & exogenous MP	400	0.7	SEM	Warning	599	1
Study 2	endogenous MP	400	0.7	SEM	Warning	588	1
Study 2	exogenous MP	100	0.3	SEM	Warning	559	3
Study 2	endogenous MP	100	0.3	SEM	Warning	305	3
Study 2	exogenous MP	100	0.3	lSAM-ULS	Warning	286	1
Study 2	endo- & exogenous MP	100	0.3	SEM	Warning	242	4
Study 2	endo- & exogenous MP	400	0.3	SEM	Warning	239	1
Study 2	no measurement MP	100	0.3	SEM	Warning	220	3
Study 2	exogenous MP	400	0.3	SEM	Warning	195	1
Study 2	exogenous MP	100	0.3	SEM	Warning	175	4
Study 2	no measurement MP	100	0.5	SEM	Warning	165	1
Study 2	endogenous MP	100	0.3	SEM	Warning	157	4

Table C2

Summary of Warnings and Errors by Condition with ID for All Studies (continued)

Study	Misspecification	N	r	Method	Type	Count	ID
Study 2	no measurement MP	100	0.3	lSAM-ULS	Warning	138	1
Study 2	no measurement MP	100	0.3	SEM	Warning	132	4
Study 2	endo- & exogenous MP	400	0.5	SEM	Warning	130	1
Study 2	endo- & exogenous MP	100	0.3	lSAM-ULS	Warning	128	1
Study 2	exogenous MP	100	0.3	lSAM-ULS	Error	121	6
Study 2	endo- & exogenous MP	100	0.5	SEM	Warning	105	2
Study 2	exogenous MP	100	0.5	SEM	Warning	93	2
Study 2	endogenous MP	100	0.3	lSAM-ULS	Warning	78	1
Study 2	exogenous MP	400	0.5	SEM	Warning	77	1
Study 2	endo- & exogenous MP	100	0.3	SEM	Warning	57	5
Study 2	no measurement MP	100	0.3	lSAM-ULS	Error	44	6
Study 2	exogenous MP	100	0.3	SEM	Warning	43	5
Study 2	endo- & exogenous MP	100	0.3	lSAM-ULS	Error	41	6
Study 2	endogenous MP	400	0.5	SEM	Warning	40	1
Study 2	endo- & exogenous MP	100	0.3	gSAM	Warning	31	4
Study 2	endogenous MP	400	0.3	SEM	Warning	29	1
Study 2	endogenous MP	100	0.3	SEM	Warning	26	5
Study 2	endogenous MP	100	0.3	gSAM	Warning	26	4
Study 2	exogenous MP	100	0.3	gSAM	Warning	20	4
Study 2	no measurement MP	100	0.3	gSAM	Warning	18	4
Study 2	endo- & exogenous MP	100	0.5	SEM	Warning	18	3
Study 2	endo- & exogenous MP	100	0.5	lSAM-ULS	Warning	17	1
Study 2	exogenous MP	100	0.5	lSAM-ULS	Warning	16	1
Study 2	no measurement MP	100	0.3	SEM	Warning	15	5
Study 2	no measurement MP	100	0.3	SEM	Warning	14	7
Study 2	exogenous MP	100	0.5	SEM	Warning	14	3
Study 2	endogenous MP	100	0.3	lSAM-ULS	Error	14	6
Study 2	endogenous MP	100	0.5	SEM	Warning	13	2
Study 2	exogenous MP	100	0.3	SEM	Warning	10	7
Study 2	exogenous MP	400	0.3	SEM	Warning	10	2
Study 2	endo- & exogenous MP	100	0.3	SEM	Warning	10	7
Study 2	endo- & exogenous MP	400	0.3	SEM	Warning	8	2
Study 2	exogenous MP	100	0.3	lSAM-ULS	Warning	7	3
Study 2	endo- & exogenous MP	100	0.3	gSAM	Error	7	6

Table C2

Summary of Warnings and Errors by Condition with ID for All Studies (continued)

Study	Misspecification	N	r	Method	Type	Count	ID
Study 2	exogenous MP	100	0.3	SEM	Warning	6	10
Study 2	exogenous MP	100	0.5	lSAM-ULS	Error	5	6
Study 2	endo- & exogenous MP	100	0.5	lSAM-ULS	Error	5	6
Study 2	exogenous MP	400	0.7	SEM	Warning	4	1
Study 2	endogenous MP	100	0.3	SEM	Warning	4	7
Study 2	no measurement MP	100	0.3	SEM	Warning	3	10
Study 2	no measurement MP	100	0.5	lSAM-ULS	Warning	3	1
Study 2	exogenous MP	100	0.3	SEM	Warning	3	11
Study 2	endo- & exogenous MP	100	0.3	SEM	Warning	3	10
Study 2	endo- & exogenous MP	100	0.3	SEM	Warning	3	11
Study 2	endo- & exogenous MP	100	0.7	SEM	Warning	3	2
Study 2	no measurement MP	100	0.7	SEM	Warning	2	1
Study 2	exogenous MP	100	0.3	gSAM	Error	2	6
Study 2	exogenous MP	100	0.3	lSAM-ULS	Warning	2	2
Study 2	exogenous MP	100	0.7	SEM	Warning	2	2
Study 2	exogenous MP	400	0.3	SEM	Warning	2	3
Study 2	endogenous MP	100	0.5	SEM	Warning	2	3
Study 2	endo- & exogenous MP	100	0.3	gSAM	Warning	2	3
Study 2	endo- & exogenous MP	100	0.5	SEM	Warning	2	4
Study 2	no measurement MP	100	0.3	gSAM	Error	1	6
Study 2	no measurement MP	400	0.3	SEM	Warning	1	1
Study 2	exogenous MP	100	0.3	SEM	Warning	1	13
Study 2	exogenous MP	100	0.3	SEM	Warning	1	12
Study 2	exogenous MP	100	0.5	lSAM-ULS	Warning	1	3
Study 2	exogenous MP	100	0.7	SEM	Warning	1	3
Study 2	exogenous MP	400	0.3	lSAM-ULS	Error	1	6
Study 2	endogenous MP	100	0.3	SEM	Warning	1	9
Study 2	endogenous MP	100	0.3	lSAM-ULS	Warning	1	3
Study 2	endogenous MP	100	0.5	lSAM-ULS	Error	1	6
Study 2	endogenous MP	100	0.5	lSAM-ULS	Warning	1	1
Study 2	endogenous MP	400	0.3	SEM	Warning	1	2
Study 2	endo- & exogenous MP	100	0.3	lSAM-ULS	Warning	1	2
Study 2	endo- & exogenous MP	100	0.3	lSAM ML	Error	1	6
Study 2	endo- & exogenous MP	100	0.5	lSAM-ULS	Warning	1	2

Table C2

Summary of Warnings and Errors by Condition with ID for All Studies (continued)

Study	Misspecification	N	r	Method	Type	Count	ID
Study 2	endo- & exogenous MP	100	0.7	lSAM-ULS	Warning	1	1
Study 2	endo- & exogenous MP	400	0.3	SEM	Warning	1	3
Study 2	endo- & exogenous MP	400	0.3	SEM	Warning	1	4
Study 3	negative CR	6400	0.5	SEM unbound	Warning	5000	1
Study 3	negative CR	400	0.5	SEM unbound	Warning	4997	1
Study 3	negative CR	6400	0.3	SEM unbound	Warning	4997	2
Study 3	negative CR	400	0.3	SEM unbound	Warning	4996	2
Study 3	negative CR	100	0.3	SEM unbound	Warning	4491	2
Study 3	negative CR	50	0.7	SEM unbound	Warning	4310	1
Study 3	negative CR	100	0.5	SEM unbound	Warning	4010	1
Study 3	negative CR	100	0.7	SEM unbound	Warning	3791	1
Study 3	positive CR	50	0.5	SEM unbound	Warning	3527	1
Study 3	positive CR	100	0.3	SEM unbound	Warning	3439	1
Study 3	negative CR	50	0.3	SEM unbound	Warning	3270	2
Study 3	negative CL	50	0.7	SEM unbound	Warning	3217	1
Study 3	positive CL	50	0.7	SEM unbound	Warning	2957	1
Study 3	negative CL	50	0.5	SEM unbound	Warning	2716	1
Study 3	positive CR	50	0.3	SEM unbound	Warning	2670	1
Study 3	negative CL	100	0.5	SEM unbound	Warning	2599	1
Study 3	positive CL	50	0.5	SEM unbound	Warning	2558	1
Study 3	negative CR	50	0.5	SEM unbound	Warning	2463	1
Study 3	negative CR	50	0.5	SEM unbound	Warning	2253	2
Study 3	negative CL	50	0.3	SEM unbound	Warning	2055	2
Study 3	positive CR	50	0.7	SEM unbound	Warning	1987	1
Study 3	positive CL	100	0.5	SEM unbound	Warning	1806	1
Study 3	positive CL	50	0.3	SEM unbound	Warning	1781	2
Study 3	negative CL	100	0.3	SEM unbound	Warning	1728	2
Study 3	negative CR	400	0.7	SEM unbound	Warning	1633	1
Study 3	negative CL	100	0.3	SEM unbound	Warning	1583	1
Study 3	negative CR	50	0.3	SEM unbound	Warning	1548	3
Study 3	negative CL	50	0.3	SEM unbound	Warning	1543	3
Study 3	positive CL	100	0.7	SEM unbound	Warning	1492	1
Study 3	positive CR	100	0.5	SEM unbound	Warning	1478	1
Study 3	positive CL	100	0.3	SEM unbound	Warning	1372	1

Table C2

Summary of Warnings and Errors by Condition with ID for All Studies (continued)

Study	Misspecification	N	r	Method	Type	Count	ID
Study 3	positive CR	50	0.3	SEM unbound	Warning	1366	3
Study 3	negative CL	400	0.3	SEM unbound	Warning	1299	1
Study 3	negative CL	100	0.7	SEM unbound	Warning	1244	1
Study 3	negative CL	50	0.5	SEM unbound	Warning	1206	2
Study 3	positive CL	50	0.3	SEM unbound	Warning	1191	3
Study 3	positive CL	100	0.3	SEM unbound	Warning	1062	2
Study 3	positive CL	50	0.3	SEM unbound	Warning	946	1
Study 3	negative CR	100	0.5	SEM unbound	Warning	891	2
Study 3	negative CL	50	0.3	SEM unbound	Warning	817	1
Study 3	positive CR	50	0.3	SEM unbound	Warning	793	2
Study 3	negative CL	100	0.3	SEM unbound	Warning	719	3
Study 3	positive CL	50	0.5	SEM unbound	Warning	688	2
Study 3	negative CL	400	0.5	SEM unbound	Warning	565	1
Study 3	negative CR	100	0.3	SEM unbound	Warning	469	3
Study 3	negative CL	50	0.5	SEM unbound	Warning	449	3
Study 3	negative CL	100	0.5	SEM unbound	Warning	424	2
Study 3	positive CL	50	0.3	SEM unbound	Warning	356	4
Study 3	positive CR	100	0.3	SEM unbound	Warning	340	3
Study 3	positive CR	100	0.7	SEM unbound	Warning	331	1
Study 3	positive CL	100	0.3	SEM unbound	Warning	329	3
Study 3	positive CL	50	0.3	SEM unbound	Warning	311	5
Study 3	positive CR	400	0.3	SEM unbound	Warning	302	1
Study 3	negative CL	50	0.7	SEM unbound	Warning	276	2
Study 3	positive CR	100	0.3	SEM unbound	Warning	244	2
Study 3	positive CL	50	0.5	SEM unbound	Warning	242	3
Study 3	positive CL	100	0.3	SEM unbound	Warning	228	4
Study 3	negative CL	50	0.3	SEM unbound	Warning	227	5
Study 3	positive CL	400	0.5	SEM unbound	Warning	225	1
Study 3	negative CL	400	0.3	SEM unbound	Warning	225	2
Study 3	positive CL	400	0.3	SEM unbound	Warning	219	1
Study 3	positive CL	50	0.3	gSAM	Warning	189	4
Study 3	negative CR	50	0.5	SEM unbound	Warning	153	3
Study 3	negative CL	50	0.3	SEM unbound	Warning	150	4
Study 3	negative CL	50	0.3	gSAM	Warning	142	4

Table C2

Summary of Warnings and Errors by Condition with ID for All Studies (continued)

Study	Misspecification	N	r	Method	Type	Count	ID
Study 3	negative CL	50	0.7	SEM unbound	Warning	129	3
Study 3	positive CR	50	0.5	SEM unbound	Warning	120	3
Study 3	positive CL	400	0.7	SEM unbound	Warning	116	1
Study 3	positive CL	100	0.5	SEM unbound	Warning	104	2
Study 3	positive CL	50	0.7	SEM unbound	Warning	101	2
Study 3	negative CR	50	0.3	SEM unbound	Warning	97	1
Study 3	negative CL	100	0.5	SEM unbound	Warning	96	3
Study 3	negative CL	100	0.3	SEM unbound	Warning	78	4
Study 3	negative CR	50	0.3	SEM bound	Warning	76	8
Study 3	positive CL	100	0.3	SEM unbound	Warning	73	5
Study 3	negative CL	50	0.3	SEM bound	Warning	73	8
Study 3	positive CL	50	0.5	SEM unbound	Warning	71	4
Study 3	positive CR	50	0.5	SEM unbound	Warning	61	2
Study 3	negative CL	100	0.3	SEM unbound	Warning	57	5
Study 3	positive CL	50	0.3	SEM bound	Warning	53	8
Study 3	negative CR	50	0.5	SEM unbound	Warning	46	7
Study 3	positive CL	50	0.5	SEM unbound	Warning	41	5
Study 3	negative CL	50	0.3	gSAM	Error	41	6
Study 3	negative CR	50	0.3	SEM unbound	Warning	39	5
Study 3	negative CR	50	0.3	SEM bound	Warning	38	14
Study 3	negative CL	400	0.3	SEM unbound	Warning	37	3
Study 3	negative CL	100	0.7	SEM unbound	Warning	36	2
Study 3	negative CR	100	0.5	SEM unbound	Warning	31	3
Study 3	positive CL	50	0.3	gSAM	Error	29	6
Study 3	negative CL	100	0.3	SEM unbound	Warning	27	7
Study 3	negative CL	50	0.5	SEM unbound	Warning	26	5
Study 3	positive CL	100	0.5	SEM unbound	Warning	25	3
Study 3	negative CR	50	0.5	SEM unbound	Warning	25	9
Study 3	positive CL	50	0.7	SEM unbound	Warning	23	3
Study 3	positive CL	100	0.3	gSAM	Warning	21	4
Study 3	negative CR	50	0.3	SEM unbound	Warning	21	7
Study 3	negative CL	50	0.3	SEM bound	Warning	19	14
Study 3	negative CR	100	0.5	SEM unbound	Warning	19	7
Study 3	negative CR	100	0.3	SEM unbound	Warning	17	7

Table C2

Summary of Warnings and Errors by Condition with ID for All Studies (continued)

Study	Misspecification	N	r	Method	Type	Count	ID
Study 3	negative CL	50	0.3	SEM unbound	Warning	16	7
Study 3	negative CL	50	0.5	SEM unbound	Warning	16	4
Study 3	negative CL	100	0.3	gSAM	Warning	16	4
Study 3	negative CR	100	0.3	SEM unbound	Warning	15	1
Study 3	positive CL	400	0.3	SEM unbound	Warning	14	2
Study 3	negative CR	50	0.7	SEM unbound	Warning	13	2
Study 3	positive CL	50	0.3	SEM unbound	Warning	12	7
Study 3	negative CR	50	0.3	SEM unbound	Warning	11	11
Study 3	positive CL	50	0.3	SEM bound	Warning	10	14
Study 3	positive CL	50	0.3	SEM bound	Warning	9	15
Study 3	positive CL	100	0.3	SEM bound	Warning	9	8
Study 3	negative CR	50	0.3	SEM bound	Warning	9	15
Study 3	negative CR	50	0.3	gSAM	Warning	9	4
Study 3	positive CL	50	0.3	SEM unbound	Warning	8	8
Study 3	positive CL	50	0.3	gSAM	Warning	8	8
Study 3	negative CR	50	0.3	SEM unbound	Warning	8	4
Study 3	negative CR	100	0.5	SEM unbound	Warning	8	9
Study 3	positive CL	50	0.3	SEM unbound	Warning	7	12
Study 3	positive CL	50	0.5	gSAM	Warning	7	4
Study 3	positive CL	100	0.7	SEM unbound	Warning	7	2
Study 3	negative CL	50	0.3	SEM bound	Warning	7	15
Study 3	negative CL	50	0.3	SEM unbound	Warning	7	8
Study 3	negative CL	50	0.3	gSAM	Warning	7	8
Study 3	negative CL	50	0.5	SEM unbound	Warning	7	7
Study 3	negative CL	100	0.3	SEM bound	Warning	7	8
Study 3	positive CL	50	0.5	SEM bound	Warning	6	8
Study 3	positive CL	100	0.3	SEM unbound	Warning	6	7
Study 3	positive CL	400	0.3	SEM unbound	Warning	6	3
Study 3	negative CL	50	0.3	SEM unbound	Warning	6	12
Study 3	negative CL	100	0.3	SEM unbound	Warning	6	11
Study 3	negative CL	100	0.3	gSAM	Error	6	6
Study 3	negative CL	100	0.7	SEM unbound	Warning	6	3
Study 3	negative CR	100	0.3	SEM bound	Warning	6	8
Study 3	negative CR	100	0.3	SEM unbound	Warning	6	9

Table C2

Summary of Warnings and Errors by Condition with ID for All Studies (continued)

Study	Misspecification	N	r	Method	Type	Count	ID
Study 3	positive CL	50	0.3	SEM unbound	Warning	5	11
Study 3	positive CL	100	0.3	gSAM	Error	5	6
Study 3	positive CL	100	0.5	SEM unbound	Warning	5	4
Study 3	negative CL	50	0.3	SEM unbound	Warning	5	11
Study 3	negative CL	50	0.5	gSAM	Warning	4	4
Study 3	negative CL	400	0.7	SEM unbound	Warning	4	1
Study 3	negative CR	100	0.3	SEM bound	Warning	4	14
Study 3	negative CR	400	0.3	SEM unbound	Warning	4	3
Study 3	positive CL	50	0.3	SEM unbound	Warning	3	10
Study 3	positive CL	50	0.3	gSAM	Warning	3	3
Study 3	positive CR	50	0.3	SEM unbound	Warning	3	5
Study 3	negative CL	50	0.3	gSAM	Warning	3	3
Study 3	negative CL	50	0.5	SEM bound	Warning	3	8
Study 3	negative CR	50	0.3	SEM unbound	Warning	3	9
Study 3	negative CR	50	0.3	gSAM	Error	3	6
Study 3	negative CR	6400	0.3	SEM unbound	Warning	3	3
Study 3	positive CL	50	0.3	SEM unbound	Warning	2	9
Study 3	positive CL	100	0.3	SEM unbound	Warning	2	10
Study 3	positive CL	100	0.3	SEM unbound	Warning	2	8
Study 3	positive CR	50	0.7	SEM unbound	Warning	2	3
Study 3	positive CR	100	0.3	SEM unbound	Warning	2	7
Study 3	positive CR	400	0.5	SEM unbound	Warning	2	1
Study 3	negative CL	50	0.3	SEM unbound	Warning	2	9
Study 3	negative CL	100	0.3	SEM unbound	Warning	2	9
Study 3	negative CL	100	0.3	gSAM	Warning	2	8
Study 3	negative CL	100	0.5	SEM unbound	Warning	2	7
Study 3	negative CR	50	0.5	SEM unbound	Warning	2	5
Study 3	negative CR	100	0.3	SEM unbound	Warning	2	11
Study 3	positive CL	50	0.3	SEM unbound	Warning	1	16
Study 3	positive CL	50	0.5	SEM unbound	Warning	1	9
Study 3	positive CL	50	0.5	SEM unbound	Warning	1	8
Study 3	positive CL	50	0.5	gSAM	Warning	1	3
Study 3	positive CL	50	0.7	SEM unbound	Warning	1	4
Study 3	positive CL	50	0.7	SEM unbound	Warning	1	5

Table C2

Summary of Warnings and Errors by Condition with ID for All Studies (continued)

Study	Misspecification	N	r	Method	Type	Count	ID
Study 3	positive CL	100	0.3	SEM unbound	Warning	1	11
Study 3	positive CL	100	0.3	gSAM	Warning	1	3
Study 3	positive CL	100	0.5	SEM unbound	Warning	1	5
Study 3	positive CR	50	0.3	SEM unbound	Warning	1	4
Study 3	positive CR	100	0.5	SEM unbound	Warning	1	3
Study 3	negative CL	50	0.3	SEM bound	Warning	1	17
Study 3	negative CL	50	0.3	SEM unbound	Warning	1	10
Study 3	negative CL	50	0.3	SEM unbound	Warning	1	18
Study 3	negative CL	50	0.5	SEM unbound	Warning	1	8
Study 3	negative CL	100	0.3	SEM bound	Warning	1	14
Study 3	negative CL	100	0.3	SEM unbound	Warning	1	8
Study 3	negative CL	100	0.5	SEM bound	Warning	1	8
Study 3	negative CL	100	0.5	SEM unbound	Warning	1	9
Study 3	negative CL	100	0.5	SEM unbound	Warning	1	4
Study 3	negative CL	100	0.5	SEM unbound	Warning	1	5
Study 3	negative CL	400	0.3	SEM unbound	Warning	1	7
Study 3	negative CR	50	0.7	SEM unbound	Warning	1	3
Study 3	negative CR	100	0.3	SEM bound	Warning	1	17
Study 3	negative CR	100	0.7	SEM unbound	Warning	1	2
Study 3	negative CR	400	0.5	SEM unbound	Warning	1	2

Note. This table summarizes the count of warnings and errors for each condition - sample size N, indicator reliability r and measurement misspecification (MP) of cross-loadings (CL) or correlated residuals (CR) - in all three simulation studies with the respective ID number corresponding to Table C1.

Appendix D: Reproducibility Instructions

All files this thesis and the simulation is based on are hosted on GitHub in the Adversarial Simulation repository. The `VK/thesis` subdirectory contains all files generating the thesis by Kriegmair using simulation study results data from the `VK/simulation/results` directory and is viewable here. The rendering of the thesis (including data visualization) is managed using Docker (Merkel, 2014) and `renv` (Ushey & Wickham, 2024) with the `VK/thesis/renv.lock` file and can be replicated here with a click on "Run workflow". The `VK/simulation/results` direcotry contains all raw and postprocessed results of the simulations conducted by Collaborator A (Kriegmair). Results and reproducability instructions of the studies conducted by Collaborator B (Kosanke) are available in the same repository in the `LK` subdirectory.

The `VK/simulation` subdirectory contains the source code and Docker setup to replicate the studies conducted in the Adversarial Simulation by Collaborator A. A simple minimal proof of reproducibility (with two repetitions) can be triggered via a designated GitHub action here, press "Run workflow" and view the results on a new pull request after about 20 Minutes. For a local replication use the following instructions:

1. Prerequisites: Ensure you have Docker installed on your system. You can download and install Docker here.

2. Clone the Repository: In your command line or terminal clone the Adversarial Simulation repository to your local machine and navigate to the simulation directory:

   ```
   git clone https://github.com/valentinkm/adversarial-simulation.git
   cd adversarial-simulation/VK/simulation
   ```

3. Run Simulation in Docker Container:

   ```
   make
   ```

   This command will run the individual studies (study 1 & study 2) as well as the "joint" study (study 3) and save the results in the `results_replic/` directory. The default number of replications is set to 2 for each study.

4. Additional Information: The R environment for the simulation is also managed using `renv` (Ushey & Wickham, 2024), and the exact package versions are recorded in the `VK/simulation/renv.lock` file.

Eigenständigkeitserklärung

Ich versichere, dass ich die vorliegende schriftliche Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe, alle Ausführungen, die anderen Schriften wörtlich oder sinngemäß entnommen wurden, kenntlich gemacht sind und die Arbeit in gleicher oder ähnlicher Form noch nicht für andere Prüfungen verwendet wurde sowie keiner anderen Prüfungsbehörde vorgelegen hat.

Ich habe alle Stellen, die dem Wortlaut oder dem Sinn nach (inkl. Übersetzungen) anderen Werken entnommen sind, unter genauer Angabe der Quelle (einschl. des World Wide Web sowie anderer elektronischer Datensammlungen) deutlich als Entlehnung kenntlich gemacht. Dies gilt auch für angefügte Zeichnungen, bildliche Darstellungen, Skizzen und dergleichen.

Zusätzlich versichere ich, dass ich beim Einsatz von IT-/KI-gestützten Werkzeugen diese Werkzeuge in der unten genannten „Übersicht verwendeter Hilfsmittel" mit ihrem Produktnamen und der Versionsnummer, meiner Bezugsquelle (z.B. URL) und Angaben zur Nutzung vollständig aufgeführt sowie die Checkliste wahrheitsgemäß ausgefüllt habe. Davon ausgenommen sind diejenigen IT-/KI-gestützten Schreibwerkzeuge, die von meinem zuständigen Prüfungsbüro bis zum Zeitpunkt der Abgabe meiner Arbeit als nicht anzeigepflichtig eingestuft wurden („Whitelist"). Bei der Erstellung dieser Arbeit habe ich durchgehend eigenständig und beim Einsatz IT-/KI-gestützter Schreibwerkzeuge steuernd gearbeitet.

Mir ist bekannt, dass bei Verstößen gegen diese Grundsätze ein Verfahren wegen Täuschungsversuchs bzw. Täuschung gemäß der fachspezifischen Prüfungsordnung und/oder der Fächerübergreifenden Satzung zur Regelung von Zulassung, Studium und Prüfung der Humboldt-Universität zu Berlin (ZSP-HU) eingeleitet wird.

Berlin, 14.11.2024 Valentin Kriegmair

……………………………………………………………………………………………..

Ort, Datum, Unterschrift

Übersicht verwendeter Hilfsmittel

Ausfüllhinweis: Bitte geben Sie an, welche der folgenden Programme Sie im Rahmen Ihrer Arbeit verwendet haben. Bitte tragen Sie dazu bei jedem Programm ein, wie Sie dieses genutzt haben. Tragen Sie dazu die jeweilige Ziffer in die Spalte "Nutzung" ein. Es können mehrere Ziffern je Programm eingetragen werden. Sollten Sie ein Pro-gramm genutzt haben, welches nicht in der Tabelle aufgeführt ist, ergänzen Sie dieses bitte unter Angabe der Nutzungsart.

Nutzung:

(0) Nicht verwendet
(1) Generierung von Ideen/Brainstorming
(2) Literaturrecherche
(3) Übersetzung von Texten
(4) Zusammenfassen von Quellen
(5) Inhalte auf andere Art und Weise erklären lassen (z. B. Konstrukte, methodische Vorgehensweisen, Analysen)
(6) Erstellen von Textabschnitten, welche als Vorlage dienen
(7) Überarbeitung von eigenen Textelementen (bitte Seitenzahl der Arbeit angeben)
(8) Auswertung von Daten (z. B. Schreiben von Code, Definieren der passenden Auswertungsmethoden, Erstellen von Abbildungen)
(9) Rechtschreibprüfung
(10) Visualisierungen zu illustrativen oder dekorativen Zwecken

Checkliste: IT-/KI-gestützte Schreibwerkzeuge:

Programm *(Bitte geben Sie auch weitere Programme an, die Sie verwendet haben, um Ihre Arbeit zu schreiben.)*	**falls relevant: betroffene Abschnitte** *(Bitte geben Sie Seitenzahlen an, sofern es nicht auf das gesamte Dokument zutrifft.)*	**Nutzung** *(Bitte Zahl eintragen – siehe oben)*
ChatGPT		8, 7, 1
Elicit.org		
Grammarly		9
GitHub Copilot		8
Perplexity AI		
ChatPDF		
DeepL		
BingChat		
Gamma		

Humanata AI		
Keenious		
Monica		
Claude 3.5 Sonnet		8

Sofern Sie sich unsicher sind, ob ein Programm unter "künstliche Intelligenz" fällt, tragen Sie es ein. Es erwachsen Ihnen keine Nachteile durch die Nennung des Programms. Sollten Sie weitere Programme verwendet haben, können Sie diese in die leeren Zeilen eintragen.

Whitelist

Folgende Programme müssen nicht aufgelistet oder bewertet werden. Diese Programme können ohne weitere Angaben genutzt werden:

- ✓ Microsoft Office, LaTeX, OpenOffice, iWork
- ✓ Google Scholar
- ✓ Datenbanken der Universitätsbibliothek
- ✓ Literaturverwaltungsprogramme (Zotero, Endnote, Mendeley, etc.)

Ggf. weitere Erklärungen: