Adversarial Simulation

A case study

Leonard Kosanke

The git hash is: 85786

Abstract

Previous research has demonstrated the potential of adversarial collaboration (AC) in empirical studies to reduce ambiguity in findings and facilitate a truthful representation of opposing viewpoints. In this study, we propose an AC framework specifically designed for Monte Carlo simulation studies to enhance methodological rigor and generalizability in this line of research. Our framework involves two parts: conducting independent preliminary studies, followed by a joint study to synthesize and debate the results, utilizing tools to structure, document, and evaluate the AC process. To test this framework, we conducted a focused case study comparing iterative structural after measurement (SAM) estimation and traditional structural equation model (SEM) estimation. After conducting our individual studies, we integrated our results. The findings revealed no substantial disagreements between the studies, and left us unable to conduct a joint study. Still, we found some preliminary evidence that AC can increase transparency, rigor, and epistemic accountability in simulation research. Furthermore, our study provides initial insights into the practical applicability and technical feasibility of AC, supported by tools like GitHub for collaborative work. Although we were unable to test our AC framework in its entirety, our findings suggest that AC holds significant potential for improving the quality of simulation studies. Future research should further explore AC in simulations to validate and extend these findings.

Introduction

Monte Carlo simulations are an extensively utilized tool for assessing and comparing statistical methods in quantitative empirical science. They are used to evaluate the performance of estimation and inference methods by analyzing them in light of known simulated population models and values. Despite the clear insight into the underlying data structures that simulations provide, they are not immune to common pitfalls that thus far have been predominantly associated with the replication crisis in empirical research (Lohmann et al. 2022). In simulation studies, a key challenge is ensuring that results can be generalized to real-world applications. These studies must ensure that the chosen performance metrics, experimental factors, and inference models accurately reflect and test real-world scenarios. Given the impracticality of simulating every possible model, method and use case, these studies inherently involve a multitude of decision-making elements potentially prone to bias. For example, the decisions of selecting evaluation criteria for competing methods (for example absolute versus relative bias of performance), selecting which models or methods to compare in the first place, and which simulation specifications to run (for example deciding which sample sizes to simulate), can have substantive impact on the results. Evidently, many issues of open science in the empirical realm are present in the context of simulation studies as well. One approach that addresses these issues at the level of value selection suggests to sample all relevant values from existing results published in the empirical literature of interest (Bollmann et al. 2015). An example application could be to decide on a simulated sample size not just based on a hunch or prior experience, but based on previous empirical studies conducted in the field. Even though this idea is fruitful and improves upon the problem of generalizability, the argument with regards to the arbitrariness of decision-making still remains. For example, deciding on which papers, or even which models within a paper to select (and which not), can still be subject to individual bias or simply chance. Additionally, conducting extensive literature reviews for all parameter values chosen is very time-consuming and difficult to apply in practice.

Another approach to tackle these challenges that has been proposed in the empirical context is Adversarial Collaboration (AC). Adversarial collaboration, piloted by Mellers, Hertwig, and Kahneman (2001) and popularized by Kahneman (2011), is a research method increasingly recognized within the quantitative empirical research community for enhancing scientific rigor. Praised as "The next science reform" adversarial collaboration is the process of

disagreeing scholars working jointly to resolve scientific disputes (Clark and Tetlock 2021). An ongoing effort by Melloni et al. (2023), for example, conducts Adversarial Collaboration between proponents of two different theories on the relationship between consciousness and brain activity, hoping to advance research in this field. Adversarial Collaboration entails identifying points of empirical disagreement, designing mutually agreed upon studies to test competing hypotheses, and jointly publishing results, irrespective of the outcome. The idea is that in conducting adversarial collaboration, fair comparison and truthful representation of opposing views can be achieved, thereby enhancing epistemic accountability and reducing research ambiguity in scientific decision-making. Additionally, juxtaposing and debating competing positions in this way can improve generalizability of results.

Our goal is to transfer the benefits of adversarial collaboration to simulation studies. To do so, we conducted a focused case study, utilizing an exemplary topic in the literature of simulation studies: comparing a newly proposed iterative structural after measurement (SAM) estimation approach to structural equation model (SEM) estimation with traditional, non-iterative SEM estimation (Dhaene and Rosseel 2023; Robitzsch 2022; Rosseel and Loh 2022). In their respective simulations, the authors' results differed up to the point of contradiction, providing us with ideal grounds for conducting adversarial collaboration. While Dhaene and Rosseel (2023) concluded that SAM estimation generally outperforms non-iterative SEM in small samples, Robitzsch (2022) did not find the methods to differ. Similarly, whereas the former found SAM to be more robust against model misspecification, the latter argued the opposite to be true. Applying Adversarial Collaboration, we, the authors of this paper, each represented one of these competing positions, as detailed later.

This allows us to examine the practical feasibility of adversarial collaboration in Monte Carlo simulation studies as well its potential to enhance methodological rigor and generalizability in this domain of research. Thus, this investigation leads us to address the following research question:

Can adversarial collaboration be applied to simulation studies, in terms of practical applicability and technical feasibility?

Methods

This section contains two parts that represent our proposed AC framework. First, I describe the procedure of the AC with its two parts and the tools it uses. Then, I propose a technical setup to facilitate its implementation.

Procedure of the Adversarial Collaboration

To answer our research question, we created and followed the following procedure for the adversarial collaboration process: The procedure consists of two parts. In the first part, each researcher conducts an independent preliminary study. In the second part, a joint study, including a joint simulation protocol, should be pursued.

Part one: Individual studies

As a first step of the individual studies, the two adversarial sides have to agree on one or more research questions that should be as specific as possible and directly tackle the disagreement. These individual studies should include the generation of individual simulation protocols, as suggested by Morris, White, and Crowther (2019). To facilitate a structured comparison and integration of studies throughout the collaboration process, we agreed on a structure for conducting our individual simulation studies in advance. This structure (see Table 1) is based on relevant literature in the field of simulation studies (Paxton et al. 2001; Morris, White, and Crowther 2019; Boomsma 2013).

Table 1

Structure of simulation studies

Step	Phase
1	**Defining Aims and Objectives**
	- Verbal description of the research question, making it specific.
	- Examples include examining goodness-of-fit statistics and comparing ML to 2SLS.
2	**Specification of Population Model**
	- **Optional** and depending on the field.
	- Modularities include:

Step	Phase
	- Structure: e.g., CFA or SEM
	- Size: number of latents and indicators
	- Complexity: cross-loaded indicators, reciprocal paths, exogenous predictors
	- Selection of target model.
3	**Data Generation Mechanism**
	- Resampling vs. parametric model draw.
	- Random number draw for data generation.
4	**Experimental Design Simulation Procedures**
	- Determine factors to vary, levels, and whether fully, partly factorial, or one at a time.
	- Examples include:
	- Sample size
	- Distribution of the observed variables
	- Extent of misspecification
5	**Method Selection**
	- Varies depending on research question.
	- Examples include type and number of estimation methods to be compared.
6	**Defining Estimands / Population Level Values**
	- Should reflect values commonly encountered in applied research.
	- Considerations for power issues and bias due to misspecification.
	- Example: Parameters should be statistically significant, even at the smallest sample size of the simulation.
7	**Performance Measures**
	- Selection and justification of use of measures such as bias, sensitivity/specificity, predictive accuracy.
	- Decision on number of simulations for acceptable Monte Carlo SE for these measures.
8	**Software Selection**
	- Software to run simulation using specific packages & functions.
9	**Analysis and Interpretation Plan**
	- Analysis: descriptive vs. inferential.

Step	Phase
	- Interpretation: decision criteria that evaluate performance (e.g., if 1-ß > 90%, the method performs well).
10	**Coding and Execution**
	- Amount and content of scripts.
	- Include (sanity) checks, setting seeds, troubleshooting & verification.
11	**Analyzing Results**
	- Descriptive, graphical, inferential, and exploration.
12	**Reporting & Presentation**
	- Provide rationales for each choice made in previous steps.
	- Publishing code and simulated data.

This structure allows for the expected divergences in results of the individual simulations, but at the same time provides a basis for systematic comparison and synthesis for the joint simulation.

Part two: Joint study

In the second part, a joint study, including a joint simulation protocol, should be pursued. This process starts based on the results of Part one. Here, the main part of the adversarial collaboration takes place. It contains multiple parts I will explain in the following sections: Each step of the individual simulation studies in Table 1 (from the substantive research question up until the interpretation of the results) is scrutinized and debated between collaborators, using adversarial collaboration techniques. Decisions are made and documented based on the most convincing argument presented, if possible. If an interest in the evaluation of the collaboration exists, diary entries can be written after each step. In the end, each collaborator reports the results in their own paper. If desired, a joint paper can be published as well.

Adversarial collaboration techniques

We are aware that adversarial collaboration has its limitations when it comes to decision making in joint studies, and has lead to unresolvable disagreements in previous studies

(Cowan et al. 2020; Mellers, Hertwig, and Kahneman 2001). In order to mitigate this risk, next to each collaborator publishing their own paper, we propose to conduct the joint study in a structured and formalized manner. To this end, we identified several collaboration techniques presented in Table 2. These were also meant to facilitate the evaluation of the adversarial collaboration with regards to our main research question.

Table 2

List of adversarial collaboration techniques

• Core Disagreements: Arrive at clearly defined core disagreements that might be the origin for conflicts. (Clark et al., 2022)
• Assumption check: List, question, and categorize assumptions (Kardos & Dexter, 2017)
• Red-Teaming: Generating what if scenarios, to identify limitations of the adversaries' approach (Kardos & Dexter, 2017)
• Quality of information (Kardos Dexter, 2017): checking the adversaries' quality of evidence based on literature.
• Third neutral arbiter: In case of fundamentally unresolvable disagreements, a third neutral arbiter (Aaron Peikert) will be consulted to try to resolve them.

Documentation - Decision log

As the number of decisions made and their documentation is large, only the most important results should be presented in the respective papers. In addition, the appendix should contain a separate decision log with a detailed and complete documentation of all decisions made. Here, the summarized results of the AC-techniques used for each step of the simulation study, as well as their consequence for the decision-making process should be detailed. Another aspect of structuring the collaboration within the decision log, lies in the way decision making is implemented in the joint study of our framework. In our mind, decision making can be based on four distinct grounds: Evidence-based, pragmatic reasons, arbitrary reasons or other reasons (e.g. personal values, political issues). Firstly, we deem a decision evidence

based, if one can find a clear answer for a disagreement, based on empirical evidence or in the literature more general. Secondly, to be able to keep the scope of this project, decisions can be made for pragmatic reasons, as for example in the presence of time constraints. Thirdly, arbitrary reasons could be any agreement where the first too grounds are not present, but still a decision has to be made. Importantly, while this reason does not help in deciding for an option directly, it helps in understanding how often there is no substantive or pragmatic reason for decisions in simulation studies. If this is the case, one might resort to deciding at random. Lastly, there might be other reasons that lead to a decision, that can not be anticipated but should still be captured.
These four reasons ground decision-making and aim to ease the collaborative process by giving a structure. In some cases, more than one of these reasons can be present. In such cases, we their relevance to a decision could be ranked, if possible. An example could be the following mock log entry:

```
Decision element: factor of sample sizes

Result: c(50, 100, 200)

Grounds: Primary - Evidence-based, as per Peikert et al. (2020);

         Secondary - pragmatic reason, average of individual suggestions.
```

Diary

The adversarial collaboration in Part 2 (Joint study) can be documented and evaluated from each collaborators subjective perspective using semi-structured diary entries after each step. We provided an exemplary diary template that includes questions based on Shah and Leeder (2016) in our github repository under this link: https://github.com/lkosanke/AdversarialSimulation/blob/main/diary-template.md. The diary has the purpose of accumulating data for the evaluation of the AC and the collaboration procedure we propose. In our case, collecting this qualitative data was supposed to help us answer our research question. The evaluation of the diary entry can be conducted in a semi-structured and comparative manner, resorting for example to analysis of quantity of words to identify common themes of the collaboration.

Technical setup

As simulation studies are computationally intensive and require a structured approach to save, share and review code, we propose using Github to ease collaboration (Peikert, Van Lissa, and Brandmaier 2021). Github allows to set up and synchronize multiple repositories to allow both individual and collaborative work. After setting up a first repository at one adversaries user account, the other party can fork that repository to have a copy at their own account. These forks can be synchronized at any time, but also left seperate. This functionality aligns well with the different parts of our procedure: For the individual studies, each adversary can work in their own repository. For the joint study, both can work in one repository, and synchronize the other one at will. Additionally, the Github releases feature allows to align milestones such as the publishing of the simulation protocols or the finished result analysis of all studies. These can even be assigned a DOI via Zenodo (European Organization For Nuclear Research and OpenAIRE 2013).

Results

The structure of this section mirrors our AC procedure. I will start presenting the preparations, methods and results of the individual studies and their substantive topics. Then, I will integrate the results of these studies to form a substantive conclusion. Lastly, I will analyze the results of the collaboration and thus answer our main research question.

Preparation of the individual studies

For our substantic topic of interest, Leonard Kosanke replicated relevant parts of the study by Robitzsch (2022), and Valentin Kriegmair replicated the study by Dhaene and Rosseel (2023).

As a first step, after reading the relevant literature in the field and informing us thoroughly on our side of the argument, we agreed on two substantive research questions to depict our disagreement as precisely as possible:

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?

Next, we started the AC by conducting our individual studies. We replicated and in some parts extended on the results of Robitzsch (2022) and Dhaene and Rosseel (2023). The next sections presents the methods and results of these studies, as well as their integration.

Individual study by Kriegmair

Methods

The following two sections contain the methods and results for the studies from Valentin Kriegmair. I added them indended, as they are direct quotes.

> The following description of the simulation studies is based on the established structure for simulation studies to coalign with the other conducted studies to facilitate a potential collaboration.
>
> **Aims, objectives and research questions**
>
> Both studies aimed to evaluate the performance of traditional SEM 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 established prior to conducting the studies served as general basis for both studies
>
> **Population Models and Data Generation Mechanisms**
>
> **Study 1:**
>
> Data were generated based on a 5-factor population structural model with 3 indicators for each factor. Four different models were simulated (see figures 1-4)
>
> - Model 1.1: Correctly specified model.
> - Model 1.2: Misspecified with cross-loadings ignored in the estimation model.
> - Model 1.3: Misspecified with wrong direction of path and omitted correlated residuals.

- Model 1.4: Misspecified with multiple omitted structural paths.

For all models, the population-level values of the structural parameters were set to 0.1. Indicator reliability levels were manipulated with factor loadings set at low ($\lambda = 0.3$), moderate ($\lambda = 0.5$), or high ($\lambda = 0.7$).

Study 2:

Data were generated based on a 5-factor population structural model with 3 indicators for each factor, similar to Study 1, but with the additional condition of varying variance explained (R^2) by the endogenous factors was set at low ($R^2 = 0.1$) or medium ($R^2 = 0.4$), so that the regression weights were either between 0.183 and 0.224 (low) or between 0.365 and 0.447 (medium). Note however that the computation of this was a simplification and did not accurately result in said R^2 values. The aim here was only generally to modulate between lower and higher regression weights. Misspecifications included omitting a residual covariance and a factor cross-loading in either the exogenous or endogenous part of the model, or both (see figures 5-6).

Experimental Design of simulation procedures

Study 1

The study varied three main conditions:

- Sample sizes: small ($N = 100$), moderate ($N = 400$), and large ($N = 6400$).
- Indicator reliability: low ($\lambda = 0.3$), moderate ($\lambda = 0.5$), high ($\lambda = 0.7$).
- Model specifications: correctly specified model and misspecified models (1.2, 1.3, and 1.4).

Study 2

The study varied five main conditions:

- Sample sizes: small ($N = 100$), medium ($N = 400$), and large ($N = 6400$).
- Variance explained by endogenous factors: low ($R^2 = 0.1$) and medium ($R^2 = 0.4$).

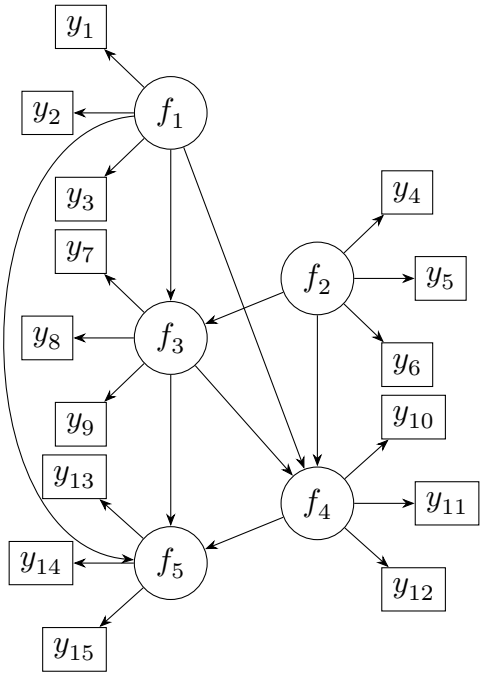

Figure 1: Note. Model 1.1: Error terms are not explicitly shown in the figure.

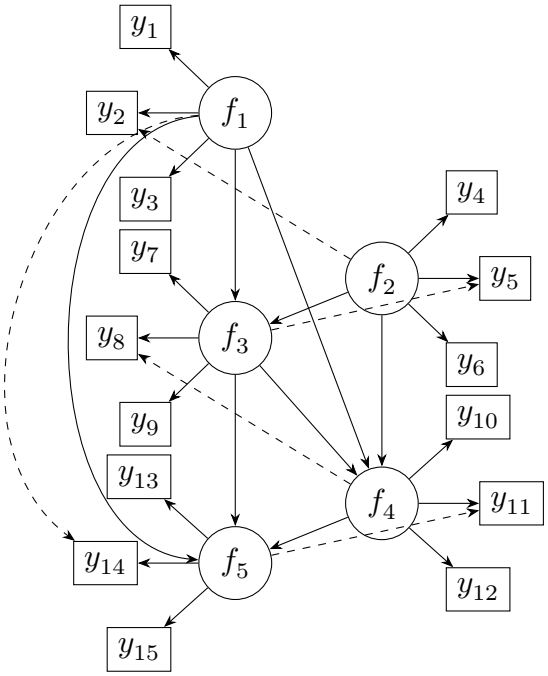

Figure 2: Note. Model 1.2: Error terms are not explicitly shown in the figure.

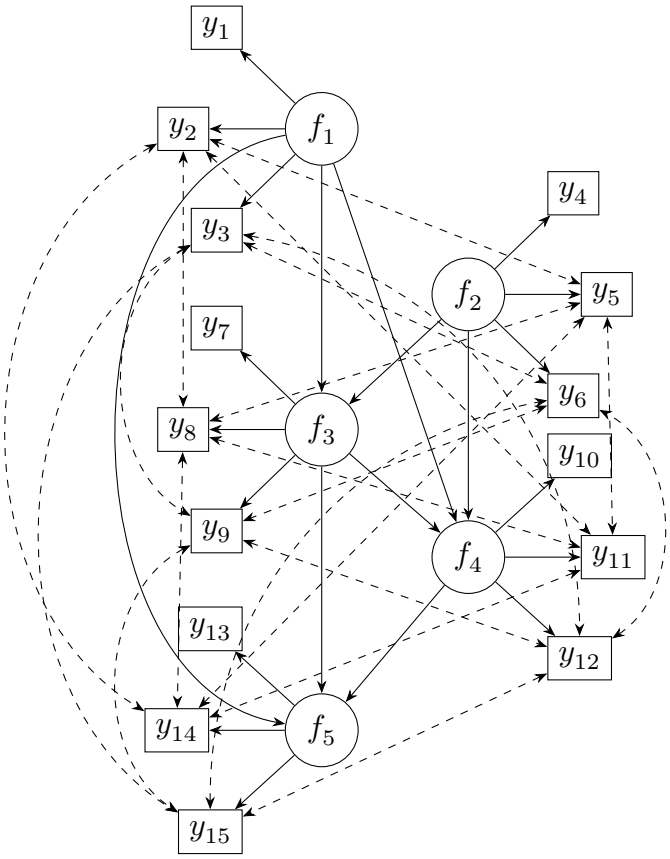

Figure 3: Note. Model 1.3: Error terms are not explicitly shown in the figure.

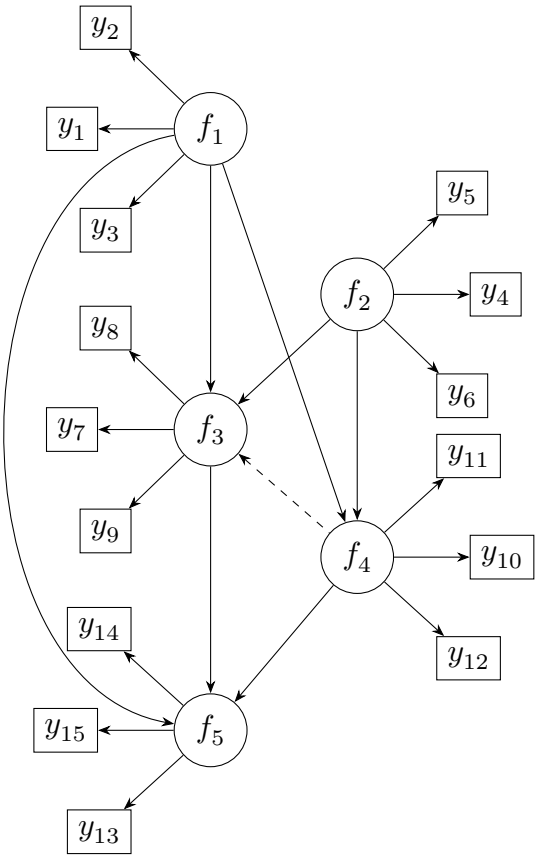

Figure 4: Note. Model 1.4: Error terms are not explicitly shown in the figure.

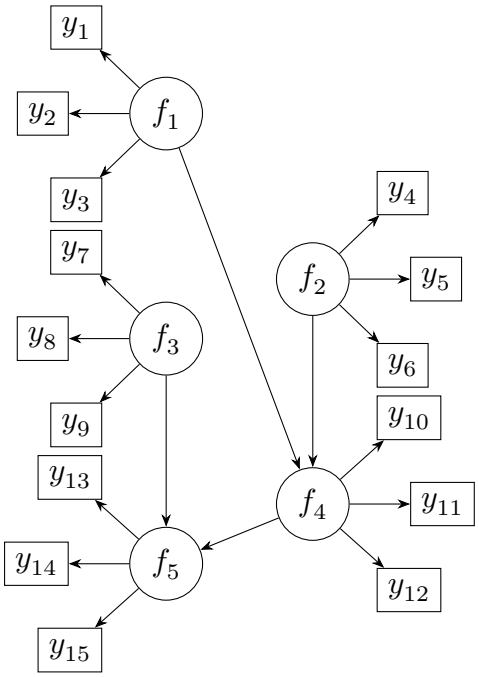

Figure 5: Note. Model 2.1: Error terms are not explicitly shown in the figure.

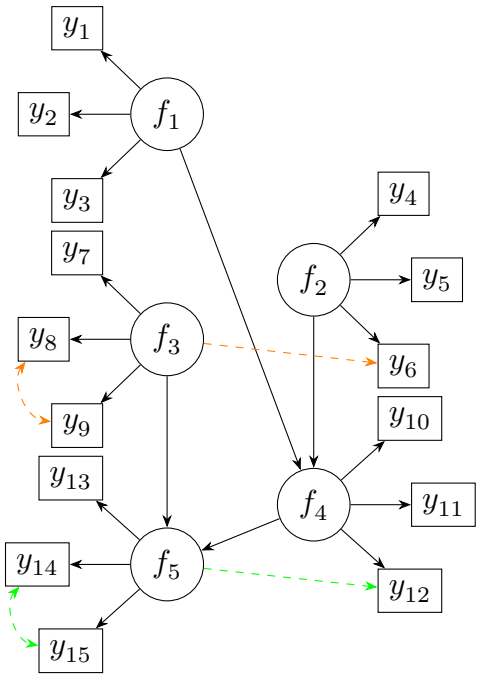

Figure 6: Note. Model 2.2: Error terms are not explicitly shown in the figure Orange lines represent misspecifications in the endogenous part of the model. Green lines represent misspecifications in the exogenous part of the model. These types of misspecifications result in different realisations of model 2.2 when they are modulated as factors in study 2 but are subsumed under one model here.

- Indicator reliability: low ($\lambda = 0.3$), moderate ($\lambda = 0.5$), and high ($\lambda = 0.7$).
- Model misspecifications: varying the population model by omitting a residual covariance and a factor cross-loading in different parts of the model.
- Number of measurement blocks: separate measurement model per latent variable ($b = 5$) and joint measurement model for all exogenous variables ($b = 3$).

Both studies were conducted with 10000 replications for each condition.

Method Selection

Study 1 and Study 2: The performance of four estimation methods were compared: Traditional SEM, global SAM (gSAM), local SAM with maximum likelihood (lSAM-ML) and local SAM with unweighted least squares (lSAM-ULS).

Performance Measures

The following performance measures were captured: Convergence rates, empirical relative biases, empirical coverage levels of 95% confidence intervals (CIs) and Root Mean Squared Errors (RMSE).

Software

All analyses were conducted in R (R Core Team 2023). See https://github.com/valentinkm/AdversarialSimulation for more details.

Analysis and Interpretation plan

The results were interpreted by descriptively comparing the performance measures (bias, MSE, convergence rates) of the different estimation methods under varying sample sizes, indicator reliability levels, and model misspecifications.

Results

As another reminder, below are the results by Valentin Kriegmair as a direct quotation. This is why they are indended.

> In the following the main patterns and trends of the results are reported and only examplary. The full results can be found in the supplementary materials.

Study 1

Convergence Rate

> For moderate and large sample sizes (N = 400, 6400), all methods achieved 100% convergence. At a small sample size (N = 100), SEM showed lower convergence rates, especially at lower reliability (0.3), while GSAM, SAM-ML, and SAM-ULS maintained high convergence rates. SEM faced significant convergence drops at the lowest reliability (0.3), particularly with misspecifications present.

Average Relative Biases

> Here this study diverged from the original of Rosseel and Loh (2022) by focusing exclusively on the average relative bias under a single type of correctly specified model (model 1), which did not include any cross loadings or correlated residuals. This simplification allowed to concentrate on the core advantage of SAM over traditional SEM: its more robust estimation of structural parameters, especially when measurement models are misspecified. For the correct model (1.1), SEM showed higher biases at smaller sample sizes and lower reliability, while GSAM, SAM-ML, and SAM-ULS exhibited lower biases. Under omitted cross loadings (model 1.2) and structural missepcifcation (model 1.4), all methods demonstrated substantial biases, particularly SEM at smaller sample sizes and lower reliability. With omitted correlated residuals and structural misspecifcation (1.3), GSAM, SAM-ML, and SAM-ULS showed more negative biases compared to SEM. Overall, GSAM, SAM-ML, and SAM-ULS were more robust with lower biases compared to SEM, especially under challenging conditions with smaller sample sizes and lower reliability. Table 1 shows average relative bias

percentages for the different estimatinon methods under varying model specifications and sample sizes in presence of omitted correlated residuals (model 1.2) and Table 2 under correlated residuals and structural misspecification.

Table 3: Average Relative Biases (in Percentages) under omitted Cross Loadings (Model 1.2)

N	Reliability	SEM	GSAM	SAM-ML	SAM-ULS
100	0.3	141.771	43.908	43.075	45.883
100	0.5	83.748	44.476	44.696	45.435
100	0.7	48.378	37.107	37.257	37.057
400	0.3	92.112	52.356	52.824	52.500
400	0.5	77.962	48.421	48.775	48.136
400	0.7	48.172	38.923	39.097	38.427
6400	0.3	81.625	54.110	54.676	53.948
6400	0.5	66.948	48.572	48.948	48.164
6400	0.7	46.667	39.142	39.316	38.585

Table 4: Average Relative Biases (in Percentages) under omitted Correlated Residuals (Model 1.3)

N	Reliability	SEM	GSAM	SAM-ML	SAM-ULS
100	0.3	-39.158	-56.090	-56.109	-55.530
100	0.5	-27.914	-32.240	-32.251	-30.610
100	0.7	-14.898	-16.446	-16.453	-15.897
400	0.3	-51.468	-53.692	-53.703	-53.543
400	0.5	-30.509	-31.317	-31.329	-31.132
400	0.7	-15.524	-15.859	-15.864	-15.711
6400	0.3	-52.334	-52.484	-52.500	-52.396
6400	0.5	-31.140	-31.225	-31.238	-31.117
6400	0.7	-15.893	-15.933	-15.938	-15.853

Coverage

For the correct model (1.1), SEM exhibited undercoverage at smaller sample sizes and lower reliability, indicating narrow CIs and less reliable estimates. GSAM, SAM-ML, and SAM-ULS showed slight overcoverage, reflecting more conservative estimates. Under the cross loadings model (1.2), all methods had substantial undercoverage, especially SEM. The correlated residuals model (1.3) consistently revealed undercoverage across all methods, with SEM showing the highest biases. For the structural model (1.4), SEM suffered from undercoverage at lower sample sizes and reliability, while GSAM, SAM-ML, and SAM-ULS performed

better with relatively consistent biases. Overall, GSAM, SAM-ML, and SAM-ULS demonstrated better empirical coverage than SEM but the overcoverage indicates too wide confidence intervals. Table 2 show the average difference between empirical coverage levels of the 95% confidence intervals (CIs) and their nominal level (95%) using each method for different reliability values and sample sizes under omitted cross loadings (model 1.2).

Table 5: Average Coverage Difference (in Percentages) under omitted Cross Loadings (Model 1.2)

N	Reliability	SEM	GSAM	SAM-ML	SAM-ULS
100	0.3	-1.008	3.671	3.650	3.451
100	0.5	-12.684	-2.277	-2.301	-3.112
100	0.7	-8.435	-4.447	-4.411	-5.164
400	0.3	-21.617	-6.320	-6.097	-6.914
400	0.5	-29.866	-16.801	-16.647	-17.457
400	0.7	-22.129	-15.936	-15.939	-16.359
6400	0.3	-65.231	-52.064	-51.036	-53.229
6400	0.5	-65.266	-54.583	-53.841	-55.944
6400	0.7	-55.541	-50.556	-50.237	-52.093

RMSE

For the correct model (1.1), SEM had higher RMSE values at smaller sample sizes and lower reliability, while GSAM, SAM-ML, and SAM-ULS demonstrated lower RMSE values. Under omitted cross loadings (1.2) and correlated residuals (1.3), SEM showed significantly higher RMSE values, although all methods exhibited increased RMSE under these conditions. The structural model (1.4) revealed that SEM had higher RMSE values at lower sample sizes and reliability, whereas GSAM, SAM-ML, and SAM-ULS showed better performance with lower RMSE. Overall, GSAM, SAM-ML, and SAM-ULS demonstrated more robustness with lower RMSE values compared to SEM, particularly in challenging conditions with smaller sample sizes and lower reliability.

Improper Solutions

Improper solutions occur when parameter estimates fall outside the boundary of the parameter space, such as when variances are estimated to be negative or

correlations exceed an absolute value of one. These issues are more likely to arise in smaller sample sizes and can indicate problems with model estimation. SEM exhibited a significantly higher count of improper solutions compared to GSAM, SAM-ML, and SAM-ULS, particularly in small sample sizes and lower reliability conditions. GSAM, SAM-ML, and SAM-ULS consistently showed minimal improper solutions across all scenarios, highlighting their robustness and reliability in parameter estimation.

Study 2

Convergence

In Study 2, all methods achieved a 100% convergence rate for moderate and large sample sizes (N = 400, 6400). For small sample sizes (N = 100), SEM showed lower convergence rates, especially at lower reliability levels (0.3), while GSAM, SAM-ML, and SAM-ULS maintained high convergence rates. R-squared values and measurement block size had minimal impact on convergence rates, indicating robustness of GSAM, SAM-ML, and SAM-ULS across various conditions, whereas SEM struggled particularly with lower reliability and smaller sample sizes.

The Influence of Number of Measurement Blocks on Bias and RMSE

For low reliability ($\lambda = 0.3$) and small sample size ($N = 100$), using 3 measurement blocks generally resulted in less negative bias compared to 5 blocks, although 5 blocks had lower RMSE. As reliability and sample sizes increased, the differences in bias and RMSE between 3 and 5 blocks diminished. Thus, for low reliability and small sample sizes, 3 measurement blocks were preferable for reducing bias, while 5 blocks performed better for RMSE. For subsequent comparisons, we considered lSAM with 3 measurement blocks as this difference was more pronounced especially in low sample size and reliability conditions.

Average Relative Bias

For the correctly specified model with low reliability ($\lambda = 0.3$) and small sample size ($N = 100$), traditional SEM showed a larger average relative bias compared to SAM methods. GSAM and lSAM-ML had small negative biases, while lSAM-ULS had a positive bias in challenging conditions. Higher R^2 had little impact on SEMs bias as absolute values, but SAM methods showed increased bias. For higher reliability and sample sizes, biases decreased for all methods.

Under misspecification (correlated residuals and crossloadings in the exogenous model), SEM had an even greater positive bias in challenging conditions, while SAM methods showed consistent negative biases, with SAM-ULS performing best.

In the endogenous misspecification model, SEM exhibited relatively small biases compared to SAM methods, which showed substantial negative biases. SEM's bias worsened with larger sample sizes and low reliability, while SAM methods' biases remained similarly large and negative.

For misspecifications in both endogenous and exogenous parts, SAM methods had consistent negative biases. SEM showed a positive bias with low reliability and sample size, turning negative with increasing sample size and higher R^2 values.

RMSE

For the correctly specified model with low reliability ($\lambda = 0.3$) and small sample size ($N = 100$), traditional SEM showed higher RMSE compared to SAM methods. GSAM and SAM-ML had lower RMSEs, while SAM-ULS had a slightly higher RMSE. Increasing regression weights (R^2) improved RMSE for both SEM and SAM methods. Under exogenous misspecifications, SEM had higher RMSE compared to SAM methods at low reliability and small sample size. This pattern persisted for endogenous misspecifications and misspecifications in both exogenous and endogenous factors, as GSAM and SAM-ML consistently showed lower RMSEs and SAM-ULS performed slightly worse. RMSE improved for all methods with higher R^2.

Improper Solutions

Table 6: Mean Relative Bias in Percentages for traditional SEM and different SAM methods under Model Misspecifications (cross loadings and correlated residuals) in Exogenous and Endogenous Factors of the Model (Model 2.2-both)

N	Reliability	R^2	SEM	GSAM	SAM-ML	SAM-ULS
100	0.3	0.1	9.39	-24.32	-16.81	-15.15
100	0.3	0.4	6.44	-25.21	-22.61	-21.45
100	0.5	0.1	5.56	-9.54	2.57	3.01
100	0.5	0.4	-4.46	-12.24	-10.20	-10.51
100	0.7	0.1	3.73	-1.85	6.55	6.26
100	0.7	0.4	-2.76	-5.81	-4.33	-5.09
400	0.3	0.1	-12.30	-22.13	-16.74	-17.62
400	0.3	0.4	-17.65	-22.40	-22.73	-23.34
400	0.5	0.1	1.59	-8.36	1.20	-0.23
400	0.5	0.4	-8.48	-11.26	-10.92	-12.13
400	0.7	0.1	2.93	-0.95	3.85	2.93
400	0.7	0.4	-3.40	-5.10	-4.56	-5.83
6400	0.3	0.1	-16.73	-21.19	-17.64	-18.85
6400	0.3	0.4	-20.17	-21.73	-23.11	-24.04
6400	0.5	0.1	-1.78	-8.08	-1.13	-2.76
6400	0.5	0.4	-9.66	-11.15	-11.28	-12.74
6400	0.7	0.1	2.17	-0.79	2.13	1.27
6400	0.7	0.4	-3.49	-5.00	-5.21	-6.97

For small sample sizes ($N = 100$) and lower reliability ($\lambda = 0.3$), SEM exhibited a high percentage of improper solutions, while GSAM, SAM-ML, and SAM-ULS consistently showed very low or zero improper solutions across all conditions.

Individual study by Kosanke

Methods

The structure of this section closely aligns to our agreed upon structure of simulation studies in Table 1.

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 3.

Table 1

Overview of the individual simulation studies by Leonard Kosanke

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

Note. CFA = Confirmatory Factor Analysis, pos. = Positive values, neg. = Negative values, RC = Residual correlation, CL = Cross-loading, N = Sample, φ = Factor correlation, β = regression weights, λ = Factor loadings.

With regards to the population models, all factors in all studies loaded onto 3 indicators each. I chose the population values to align to the original paper by Robitzsch (2022). For more details on the exact values of each study, see the simulation scripts in the github repository. The multivariate normally distributed data was generated parametrically, based on a specified population model. All simulations were conducted using seeds to allow for reproducibility.

Experimental Design of simulation procedures

Overall, 3 different types of factors were varied that can be deducted from Table 3 and are detailed again in the simulation scripts provided. Firstly, we 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. In Studies 1b and 4a, I varied some population values for model parameters (phi,

beta and/ or lambda), as well. 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 looked at 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 and 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 maximum likelihood and unweighted least squares estimation, so that loadings and variance parameters were given the constraints that they had to be positive and larger than 0.01. These were implemented in classical SEM-estimation, as well as both in their LSAM and GSAM variants. 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, Ulitzsch, and Robitzsch 2021; Ulitzsch, Lüdtke, and Robitzsch 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. I used the packages lavaan, purrr, tidyverse, furrr to conduct the simulations, as well as knitr and kableExtra for presenting the results (Rosseel 2012; Wickham and Henry 2023; Wickham et al. 2019; Vaughan and Dancho 2022; 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.

Results

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

Mainly Studies 1 and 4 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 4, only slight, but no substantial differences could be observed in terms of bias and RMSE between the LSAM- and classical ML-estimation. This is visible in Table 4 for the

example of RMSE. Here, LSAM-ML appeared to outperform both SEM-estimation methods only in N=50. For higher N, no differences were present between any of the estimators.

Table 4

Study 4a: RMSE for DGM 1.

	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. Condition: DGM 1

In Study 1, both traditional SEM approaches outperformed all SAM estimators in small to moderate samples of N=50-500. This was true for both relative bias and RMSE, and visible for the former in Table 5. Here, SAM's negative small sample bias is already visible as well.

Table 5

Study 1: Relative bias in conditions without misspecification.

	Sample Size						
Method/Metric	**50**	**100**	**250**	**500**	**1000**	**2500**	**100000**
SEM_ML	-0.045	-0.011	0.001	-0.003	0.003	-0.002	-0.000
SEM_ULS	0.024	0.022	0.012	0.002	0.006	-0.001	-0.000
LSAM_ML	-0.394	-0.270	-0.111	-0.056	-0.022	-0.011	-0.000
LSAM_ULS	-0.393	-0.270	-0.111	-0.056	-0.022	-0.011	-0.000
GSAM_ML	-0.394	-0.270	-0.111	-0.056	-0.022	-0.011	-0.000
GSAM_ULS	-0.393	-0.270	-0.111	-0.056	-0.022	-0.011	-0.000

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 6, both SEM estimators outperformed all four SAM estimators in terms of relative bias with two negative residual correlations present.

Table 6

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

	Sample Size						
Method/Metric	**50**	**100**	**250**	**500**	**1000**	**2500**	**100000**
SEM_ML	-0.205	-0.175	-0.166	-0.168	-0.161	-0.166	-0.164
SEM_ULS	-0.139	-0.145	-0.159	-0.167	-0.163	-0.170	-0.169
LSAM_ML	-0.498	-0.385	-0.272	-0.225	-0.196	-0.189	-0.180
LSAM_ULS	-0.497	-0.385	-0.272	-0.225	-0.196	-0.189	-0.180
GSAM_ML	-0.497	-0.385	-0.272	-0.225	-0.196	-0.189	-0.180
GSAM_ULS	-0.496	-0.385	-0.272	-0.225	-0.196	-0.189	-0.180

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.

Importantly, no differences between the two approaches arose in these two examples in terms of RMSE, as can be seen in Table 7 for the negative cross-loadings in study 2.

Table 7

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

	Sample Size						
Method/Metric	**50**	**100**	**250**	**500**	**1000**	**2500**	**100000**
SEM_ML_rmse	0.480	0.382	0.257	0.211	0.182	0.172	0.161
SEM_ULS_rmse	0.477	0.367	0.241	0.205	0.182	0.175	0.166
LSAM_ML_rmse	0.486	0.421	0.323	0.269	0.232	0.216	0.201
LSAM_ULS_rmse	0.487	0.421	0.323	0.269	0.232	0.216	0.201
GSAM_ML_rmse	0.490	0.422	0.323	0.269	0.232	0.216	0.201
GSAM_ULS_rmse	0.490	0.421	0.323	0.269	0.232	0.216	0.201

Conditions with positively valenced unmodelled parameters

In terms of performance for positively valenced cross-loadings and residual correlations, SAM appeared to slightly outperform traditional SEM estimation, but not in all scenarios.

Table 8 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 8

Study 3: 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

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 9 for DGM 2 of Study 4. In other conditions, however, no substantial differences arose in terms of RMSE.

Table 9

Study 4: 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

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 10 shows this was especially apparent in a sample size of 50.

Table 10

Study 1b: Absolute Bias of LSAM-ML for N=50

	Phi values				
Lambda	**phi=0**	**phi=0.2**	**phi=0.4**	**phi=0.6**	**phi=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 that absolute values of bias were calculated in my paper. 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, 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 11 shows the performance of SEM-ML, whereas Table 12 shows the performance of LSAM-ML in DGM 2 (in presence of unmodelled cross-loadings).

Table 11

Study 4a: Absolute bias of SEM-ML for DGM 2

	Sample size						
beta	**50**	**100**	**250**	**500**	**1000**	**2500**	**100.000**
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

Table 12

Study 4a: Absolute bias of LSAM-ML for DGM 2

	Sample size						
beta	**50**	**100**	**250**	**500**	**1000**	**2500**	**100.000**
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
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

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.

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.

Conclusions from both studies

Comparing the result from both individual studies, it became evident that they are at no point contradictory or in disagreement. This is because of some central distinctions between the studies. Firstly, same as in Robitzsch (2022), the use of constrained ML estimation solves both the convergence issues in small samples, as well as the positive small sample bias of the unconstrained version of the estimator. Additionally, the observed performance differences disappeared, when accounting for SAM's negative small sample bias.

Secondly, the results with regards to negatively valenced unmodelled residual correlations and cross-loading paint a more complete picture of the comparative performance of the two estimation methods. Unlike Kosanke, Kriegmair did not include these conditions. The apparent performance advantages for traditional SEM in these conditions balanced the advantages of SAM in positive conditions. In the context of empirical studies, the valence of potential residual correlations and cross-loadings is not known a priori. If that would be the case, one could simply turn to an estimator that is biased in the appropriate direction, to compensate this misspecification. Thus, achieving unbiased estimates is desirable in application contexts.

The evident performance differences observed between traditional SEM and SAM estimation in the study by Kriegmair are thus to be interpreted differently for all effects, but align with the results by Kosanke, in this light:

This is true for all performance measures, even for coverage, which was not investigated in the study by Kosanke. Coverage is a measure that is partly determined by bias. Results in the studies by Kriegmair show, that SAM showed overcoverage due to its wider confidence intervals. Consequently, SAM was less biased but also less accurate than traditional SEM. Thus, SAM only partly outperformed tradional SEM, even in the unconstrained version and the results with regards to coverage do not change our conclusions. Kriegmair's results also hinted at the present of SAM's negative small sample bias, as the largest performance differences were observed in conditions of low reliability, i.e lambdas, where the bias was especially pronounced (TABLES???). Moreover, in Kriegmairs' study 1, SAM performed even worse than unconstrained SEM for conditions of low reliability and sample size (TABLE???).

One set of findings, that hinted towards the possibility of SAM outperforming SEM was the slightly stronger robustness of SAM over constrained SEM for misspecifications along higher values of phi in Kosanke's Study 4a. Only traditional SEM seemed to be negatively affected

for higher beta values, compared to lower beta values in conditions with unmodelled cross-loadings (DGM 2). This showed slightly, but not substantially in conditions with residual correlations as well (DGM3), in the same study by Kosanke. Kriegmair's study 2 showed this finding to not be consistent in the end, as in his conditions for model 2.2 in the endogenous conditions, for low sample size and reliability, the opposite trend was visible. Here, SEM performed better, even in the unconstrained version in the presence of both cross-loadings and residual correlations (SEE TABLE???). Additionally, having in mind the differential effects for conditions of negative unmodelled parameters as well, we do not assume SAM to be consistently preferrable in this specific context.

Importantly, besides the use of coverage as a performance measure, other differences in the design and implementations were present when comparing our respective studies. These have to be adressed:

Unlike Kosanke, Kriegmair investigated additional effects with regards to endo- and exogeneity of predictors. Results showed, that the positive Bias of unconstrained SEM is stronger, if misspecification is present in exogenous predictors, if compared to it being present in endogenous and both endo- and exogenous predictors. In these latter conditions, SAM performed even worse. At the end of the day, all this finding does is limiting the negative impacts of using unconstrained SEM to an even more specific set of model misspecifications. Thus, unconstrained SEM can be viewed in an even better light relatively, for the endogenous as well as both endo- and exogenous predictor misspecifications. This does not warrant further investigation for our research question.

Furthermore, only Kriegmair investigated the differential effects of using a different number of measurement blocks in LSAM estimation. Results were mixed, as for Bias, using less measurement blocks appeared to lead to better performance, while the opposite was true in terms of RMSE. In consequence, no clear effect, that would suggest a differential performance of LSAM was present. We concluded that these findings have no relevance for our investigations on the relative comparison of SAM vs. SEM.

The difference of observing effects of regression weigths directly in the study by Kosanke, and total variance explained by Kriegmair, are also minor, as both are aimed at modifying the regression weigths and lead to comparable parameter values. For Kosanke's studies 4 and 4a, they were fixed between 0.1-0.4, while they varied between 0.183 and 0.447 in Kriegmair's study 2.

Another difference is the inclusion of conditions that examine the misspecification of the regression directionality (i.e. switching predictor and outcome variable incorrectly in the estimation model) in Kriegmair's studies. Important to mention here is that we did not observe any clear differences depending on the type of misspecification looking across all studies, and Kriegmairs studies in specific. As the same was true even in the original papers for our estimators of interest, we saw no reason to suspect that results should be different if looking at constrained ML and negative misspecified paramer values for misspecification of the regression directionality (Dhaene and Rosseel 2023; Robitzsch 2022; Rosseel and Loh 2022). Here, the same argument as before can be made: Only unconstrained ML-estimation and positive values of misspecified parameters were investigated. The observed effects should disappear when accounting for this as well as the negative small sample bias of SAM.

The results of the individual studies leave us with no imminent call for additional investigation in the joint study. What we and others viewed as a disagreement up to the point of contradiction, did not turn out to be one, upon closer examination. In consequence, we did not conduct a joint study and stopped the collaboration at this point.

Results of the collaboration

Going back to the outline of our collaboration framework, we successfully agreed on two research questions, conducted individual studies and integrated their results. Consequently, we did not find remaining disagreement or additional scenarios of interest that warranted the implementation of the second part of our AC procedure, the joint study. This means that we were unable to completely test our framework. Still, we have gained some insights with regards to our research question of whether adversarial collaboration can be applied to simulation studies, in terms of practical applicability and technical feasibility.

Practical applicability

Already the first step of agreeing on a joint research questions is a collaborative effort that requires to change perspective, and more deeply understand what the adversaries did in their previous studies. We see the main success of the research question formulation in the focus on small to moderate sample sizes and on a limited number of estimators. This allowed us to focus on specific parts of our replications and lead me, for example, to include an

additional sample size in all my studies, as well as omit studies 5 and 6 that were conducted on the population level in the paper by Robitzsch (2022). For the same reasons, Kriegmair focussed on less correctly specified conditions than the original paper in his study 1 (Rosseel and Loh 2022). Additionally, our general focus was much narrower because of this as well, as we did not include other discussions in the original papers, as for example on whether two-step-estimation is generally desirable as it seperates the definition from the measurement of latent variables (Robitzsch 2022).

When conducting the individual studies, we knew that we would have would have to be completely transparent with our results and had to give explanations for our decisions, as they would be scrutinized by the adversary later on. This increased quality of reporting, rigor and transparency in our individual studies. After the individual studies, we integrated our results and looked for disagreement and points for further investigation. We achieved a fair and truthful comparison of views and could agree that they were not opposing each other. We also believe that we increased generalizability simply by integrating the findings of two studies into one conclusion for more data generating scenarios and conditions than one individual study conducted.

With regards to the practical applicability of the collaboration of the second part, we want to emphasize that the reason for not conducting the second part was not that it was not practically applicable, but that there was no substantial reason to do it. Applying our framework, we observed the possibility to converge on research questions and found our individual simulations to be generally very compatible. This can be seen as preliminary evidence, that the second part of the collaboration should be practically applicable as well. Still, at the end of the day we could not conduct the second part of our collaboration framework. We believe that here, a lot of potential lies to further promote our intended goals and investigate our research question.

Technical feasibility

Technical feasibility in the collaboration was achieved for the parts we could implement in our case study. Utilizing tools like github with its functionalities to review code and jointly work on the same project allowed us to conduct simulations collaboratively in terms of comparing results, giving feedback and having access to the adversaries individual studies within the same github repository.

Next to these aspects, which we integrated in our AC framework, two other aspect of technical feasibility were present for us while conducting our individual simulation studies. Firstly, we had access to a computing cluster. This allowed a higher limit on computational complexity, and thus more freedom to include additional conditions to test, without fearing too much runtime or having to limit the number of repetitions. Additionally, feasibility was facilitated by the use of the furrr package in R, which allowed for parallelisation of the code and significantly reduced the runtime of our simulations (Vaughan and Dancho 2022). Again, as we could not conduct the second part of our collaboration framework, the technical feasibility of the collaborative study can not be judged conclusively.

Discussion

In this focused case study, we aimed to conduct adversarial collaboration to investigate the question if adversarial collaboration can be applied to simulation studies, in terms of practical applicability and technical feasibility. To do so, we created and tested a collaboration framework that start with each adversary conducting an individual study, before collaborating in a joint study. For this, we agreed on two substantial research questions to judge the comparative performance of traditional SEM vs. SAM estimation in small to moderate samples and under misspecifications. After finishing our individual studies, the supposed disagreement turned out not to be one, as results aligned well. Thus, we did not conduct a joint study and were only partly able to answer our research question.

For the first part of our framework, we found some evidence for practical applicability and technical feasibility: We were able to agree on two focused research question, that nudged our individual studies to focus on points of potential disagreement. Additionally, we were both able to successfully plan and conduct our individual simulations in a way that allowed integration of our results and to converge on a shared understanding of their consequences. Thus, we achieved a fair and truthful representation of both sides views. Technical feasibility was present with the use of github for sharing and accessing our own and the adversaries results. Additionally, usage of a computing cluster and parallelisation allowed us to freely include conditions where we deemed it necessary, in our individual studies.

For the second part of our framework, it remains unclear whether collaboration would be applicable and technically feasible. Another attempt should be pursued to be able to answer

our research question in the second part of our AC framework.

Even though we did not get to the point of direct collaboration in the second part of our framework, we are confident that it should be practically applicable and technically feasible: Facilitating advancement in discussion between adversaries with the structured use of the tools we presented, we are confident that the necessary convergence for a joint study is achievable. This should be even more so the case, because of some key differences between empirical and simulation studies: Firstly, opposing opinions in simulation studies are less incomensurable in terms of testing. In empirical studies, opposing theories might have completely different paradigms, measures and even construct definitions to answer the same research question. In simulation studies, these differences are much smaller. There is much more consensus on what data generating scenarios of interest could be, how performance is measured, and what the opposing methods are. Another difference is the luxury of simulation studies to have much fewer resource restrictions. The main resources are time and computation power, which usually are not too limited. This allows adversaries to not have to find compromise in all cases, and, if in doubt, just include all conditions that might be relevant. In empirical studies, there are much more limited resources such as number of participants, money, labs and much more. Thus, aspects of power have to always be considered and this leads to having to have much more compromise in AC.

In terms of *technical feasibility*, we believe that using tools like github with its functionalities to review code and jointly work on the same project, should allow to conduct simulations collaboratively. We see a lot of potential for these tools to be used in the second part of our collaboration framework, as they allow to split work within one study, give feedback on each others implementations via pull-requests, and much more. Ideally, additional resources like a computing cluster should be present to guarantee a higher limit on computational load, and thus more freedom to include additional conditions to test. This, however, should not be a problem for research conducted in universities or institutes that usually have access to computing clusters. Still, besides these assumptions and learnings from our first case study, we were unable to test the main part of collaboration and can not conclude that the practical applicability and technical feasibility is completely given. We believe that in the second part of our framework lies potential lies to further promote our intended goals and investigate our research question.

With regards to limitations there are several other aspects to consider. We are aware that we tested a very superficial "simulation" of adversariality. Neither side had long scientific

experience in our substantial field of interest, and thus the stakes were much lower than as if the original authors would have collaborated. In that case, collaboration most likely would have been more difficult. Still, including the individual studies as a first step, we think we did a good job of priming ourselves to the views of our respective sides of the argument.

Another important aspect to discuss is whether we should have been able to see the lack of disagreement earlier, even before conducting our individual studies. As most of our findings were replications of the original papers, we might have been able to see their alignment if our preparation of the substantive topic would have been more in depth. To avoid this, collaborators should invest enough time and attention to detail into the evaluation of all the papers of interest before deciding on adversarial collaboration, to make sure that there is actually conflicting evidence present.

Another question is in how far the advantages we observed with regards to increased rigor, transparency and open access were incremental benefits of adversarial collaboration, over and above the usage of simulation protocols and our open access approach to science. We believe this not to be problematic, as AC and other aspects of open science can complement each very well and reinforce positive effects. Thus, researchers who are not generally interested in open science, but in collaboration, would still be nudged to work more openly, which we view as a positive outcome.

In the end, we believe that it would be worthwhile for future research to conduct another case study like ours. Science is, in the end, a collaborative effort. Exploring the limits of collaboration is thus valuable to learn more about opportunities to advance scientific progress.

Bibliography

Bollmann, Stella, Moritz Heene, Helmut Küchenhoff, and Markus Bühner. 2015. "What Can the Real World Do for Simulation Studies? A Comparison of Exploratory Methods." https://doi.org/10.5282/UBM/EPUB.24518.

Boomsma, Anne. 2013. "Reporting Monte Carlo Studies in Structural Equation Modeling." *Structural Equation Modeling: A Multidisciplinary Journal* 20 (3): 518–40. https://doi.org/10.1080/10705511.2013.797839.

Clark, Cory, and Philip Tetlock. 2021. "Adversarial Collaboration: The Next Science Reform." In. https://doi.org/10.1007/978-3-031-29148-7_32.

Cowan, Nelson, Clément Belletier, Jason M. Doherty, Agnieszka J. Jaroslawska, Stephen Rhodes, Alicia Forsberg, Moshe Naveh-Benjamin, Pierre Barrouillet, Valérie Camos, and Robert H. Logie. 2020. "How Do Scientific Views Change? Notes From an Extended Adversarial Collaboration." *Perspectives on Psychological Science* 15 (4): 1011–25. https://doi.org/10.1177/1745691620906415.

Dhaene, Sara, and Yves Rosseel. 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): 926–40. https://doi.org/10.1080/10705511.2023.2220135.

European Organization For Nuclear Research, and OpenAIRE. 2013. "Zenodo." CERN. https://doi.org/10.25495/7GXK-RD71.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. Thinking, Fast and Slow. New York, NY, US: Farrar, Straus; Giroux.

Lohmann, Anna, Oscar L. O. Astivia, Tim P. Morris, and Rolf H. H. Groenwold. 2022. "It's Time! Ten Reasons to Start Replicating Simulation Studies." *Frontiers in Epidemiology* 2 (September): 973470. https://doi.org/10.3389/fepid.2022.973470.

Lüdtke, Oliver, Esther Ulitzsch, and Alexander Robitzsch. 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 (April). https://doi.org/10.3389/fpsyg.2021.615162.

Mellers, Barbara, Ralph Hertwig, and Daniel Kahneman. 2001. "Do Frequency Representations Eliminate Conjunction Effects? An Exercise in Adversarial Collaboration." *Psychological Science* 12 (4): 269–75. https://doi.org/10.1111/1467-9280.00350.

Melloni, Lucia, Liad Mudrik, Michael Pitts, Katarina Bendtz, Oscar Ferrante, Urszula

Gorska, Rony Hirschhorn, et al. 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.

Morris, Tim P., Ian R. White, and Michael J. Crowther. 2019. "Using Simulation Studies to Evaluate Statistical Methods." *Statistics in Medicine* 38 (11): 2074–2102. https://doi.org/10.1002/sim.8086.

Paxton, Pamela, Patrick J. Curran, Kenneth A. Bollen, Jim Kirby, and Feinian Chen. 2001. "Monte Carlo Experiments: Design and Implementation." *Structural Equation Modeling: A Multidisciplinary Journal* 8 (2): 287–312. https://doi.org/10.1207/S15328007SEM0802_7.

Peikert, Aaron, Caspar J. Van Lissa, and Andreas M. Brandmaier. 2021. "Reproducible Research in R: A Tutorial on How to Do the Same Thing More Than Once." *Psych* 3 (4): 836–67. https://doi.org/10.3390/psych3040053.

R Core Team. 2023. "R: A Language and Environment for Statistical Computing." Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robitzsch, Alexander. 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): 631–72. https://doi.org/10.3390/stats5030039.

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

Rosseel, Yves, and Wen Wei Loh. 2022. "A Structural After Measurement Approach to Structural Equation Modeling." *Psychological Methods*, No Pagination Specified–. https://doi.org/10.1037/met0000503.

Shah, Chirag, and Chris Leeder. 2016. "Exploring Collaborative Work Among Graduate Students Through the C5 Model of Collaboration: A Diary Study." *Journal of Information Science* 42 (5): 609–29. https://doi.org/10.1177/0165551515603322.

Ulitzsch, Esther, Oliver Lüdtke, and Alexander Robitzsch. 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): 527–57. https://doi.org/10.1037/met0000435.

Vaughan, Davis, and Matt Dancho. 2022. *Furrr: Apply Mapping Functions in Parallel Using Futures*.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino Mc-

Gowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, and Lionel Henry. 2023. *Purrr: Functional Programming Tools*. https://purrr.tidyverse.org/.

Xie, Yihui. 2024. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zhu, Hao, Thomas Travison, Timothy Tsai, Will Beasley, Yihui Xie, GuangChuang Yu, Stéphane Laurent, et al. 2024. "kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax." https://cran.r-project.org/web/packages/kableExtra/index.html.