Adversarial Simulation

A case study

Leonard Kosanke

The git hash is: 5fb60

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 on which sample size 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. First, an outline of the AC framework we agreed upon before the start of the collaboration. Then, each sides describes the methods they used to answer the substantive research question of the individual studies.

Outline of the Adversarial Collaboration framework

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 as a replication of their side of the argument. In our case, Leonard Kosanke replicated relevant parts of the study by Robitzsch (2022), and Valentin Kriegmair replicated the study by Dhaene and Rosseel (2023). The replications should include the generation of individual simulation protocols, as suggested by Morris, White, and Crowther (2019). In the second part, a joint study, including a joint simulation protocol, should be pursued. Here, the main part of the adversarial collaboration takes place. Each step of the individual simulation studies (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. This process starts based on the results of Part 1. To facilitate a structured comparison and integration of studies through the collaboration process, we agreed on a framework for conducting our individual simulation studies in advance, that is visible in Table 1.

Table 1: Structure of simulation studies 1. Defining aims and objectives / Research Question of Interest. (= verbal description) o as specific as possible o e.g. examination of goodness-

of-fit statistics under varying degrees of misspecification o comparison of the maximum likelihood (ML) to (2SLS) 2. Specification of Population Model o Optional and depending on field of simulation Modularities: o Structure: (e.g. CFA or SEM) o Size: number of latents and indicators o Complexity: (cross-loaded indicators, Reciprocal paths, Exogenous predictors) → select target model for (assumed) general model types 3. Data Generation Mechanism o resampling vs. parametric model draw o random number draw for data generation 4. Experimental Design Simulation Procedures o Determine what factors to vary, on what levels, and whether fully, or partly factorially or one at a time (factor = scenario in Morris et al., 2019) o e.g. sample size distribution of the observed variables extent of misspecification … 5. Method Selection o Varies depending on research question o e.g. type of and number of estimation methods to be compared 6. Defining Estimands / Population level values o should reflect values commonly encountered in applied research. o e.g. R2 values the chosen coefficients produce should also be reasonable for applied research. parameters of the model should be statistically significant, even at the smallest sample size of the simulation. consider power issues (e.g.:enough power to detect misspecification, too much power to detect misspecification at all sample sizes?) "bias" in the estimates that will be introduced by the misspecification. 7. Performance Measures o Selection and justification of use of e.g. Bias sensitivity/ specificity predictive accuracy o decision on number of simulations for acceptable Monte Carlo SE for these measures 8. Software Selection o to run simulation o packages & functions 9. Analysis and Interpretation plan o Analysis: descriptive vs. inferential o Interpretation: decision criteria that evaluate performance : e.g. if 1-ß $>$ 90, the method performs well

Before Coding and Execution: Anticipating all critical decision processes (e.g. exclusion criteria for "imperfect" samples) 10. Coding and Execution o Amount and content of scripts o include (sanity) checks o setting seeds o troubleshooting & verification 11. Analyzing results o descriptive o graphical o inferential o exploration 12. Reporting & Presentation o provide rationales for each choice made in the previous steps o publishing code and simulated data

End of Table 1

This framework 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. In our case study, we have identified two substantive research questions to co-align the individual simulation studies:

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?

In the end, each collaborator reports the results in their own paper.

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: Adversarial Collaboration (AC) Techniques:

At each step of decision making in the Joint Study, any of the following techniques

Thesis of: Leonard Kosanke

can be used, depending on their applicability:

• 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.

End of table 2

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 is the log entry in figure 1:

Figure 1: Exemplary log entry for deciding for a range of sample sizes

Decision element: factor of sample sizes

Result: c(50, 100, 200)

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

Secondary – pragmatic reason, based on the average of individual suggestions.

Evaluation - Diary entries

End of figure 1

Diary

The adversarial collaboration in Part 2 (Joint study) should be documented and evaluated from each collaborators subjective perspective using semi-structured diary entries based on Shah and Leeder (2016), after each step. This 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.

Methods for Replicating Robitzsch (2022)

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 substantive research questions:

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?

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 1.

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 1 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 are present in this study. Additionally, the amount of misspecification was varied in all studies, either via different numbers of unmodelled residual correlations, cross-loadings, or both. In Studies 1b and 4a, some population values for model parameters (phi, beta and/ or lambda) were varied, as well. Besides studies 1 and 2, full factorial designs were implemented. 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 do not give reasoning for this decision

themselves. Studies 4 and 4a 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). 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 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.

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 showed already in other simulations, using constrained maximum likelihood estimation should resolve convergence issues of classical maximum likelihood estimation in smaller samples. 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.

For interpretation of results, I oriented on cut-offs that were used in the original paper by Robitzsch (2022). For Bias, I interpreted difference 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.

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.

Results

Adversarial collaboration

Our simulation protocols can be accessed in our Github-repository, under releases: https://github.com/lkosanke/AdversarialSimulation/tree/main/.

Individual study by Leonard Kosanke

The full result analysis is available here: https://github.com/lkosanke/AdversarialSimulation/blob/main/LK/results.pdf. Here, I will focus on the most important results only. For the most part, results from Robitzsch (2022) have been successfully replicated. With regards to all studies conducted, 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.

Comparative performance in 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 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 1 for the example of RMSE. Here, LSAM-ML appeared to slightly outperform constrained SEM-ML estimation only in N=50.

Condition:

	Sample Size						
Method/Metric	**50**	**100**	**250**	**500**	**1000**	**2500**	**1e+05**
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_0

In Study 1, both traditional SEM approaches outperformed all SAM estimators in small to moderate samples of N=50-500. This is true for both relative bias and RMSE, and visible for the former in Table 2. Here, SAM's small sample bias is already visible as well.

Condition:							
	Sample Size						
Method	**50**	**100**	**250**	**500**	**1000**	**2500**	**1e+05**
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

Note:

Condition: 0_0.12

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

Condition:							
	Sample Size						
Method	**50**	**100**	**250**	**500**	**1000**	**2500**	**1e+05**
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

Note:

Condition: 2_-0.12

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 X for the negative cross-loadings in study 2.

Condition:

	Sample Size						
Method	**50**	**100**	**250**	**500**	**1000**	**2500**	**1e+05**
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

Note:

Condition: 2_-0.3

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 3 shows this finding in Study 3, in the conditions with both one unmodelled residual correlation and cross-loading, only from N=100-1000.

Study 3:							
	Sample Size						
Method/Metric	**50**	**100**	**250**	**500**	**1000**	**2500**	**1e+05**
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, only for smaller samples.

With regards to RMSE, results were mixed as well. LSAM appeared to outperform in Table X for DGM 2 of Study 4. In other conditions, however, no substantial differences arose in terms of RMSE.

```
create_styled_table(rmse_s4[["DGM_1"]], "DGM_1")
```

Condition:							
	Sample Size						
Method/Metric	**50**	**100**	**250**	**500**	**1000**	**2500**	**1e+05**
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:

Condition: DGM_1

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 5 showed this to be especially apparent in a sample size of 50.

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. Comparing LSAM-ML and -ULS estimation, results were very similar, overall. 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 remains relevant for low values of phi, unlike in the original paper by Robitzsch (2022). In consequence, there are no conditions were SAM's small sample bias is negligible.

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 X shows the performance of SEM-ML, whereas Table Y shows the performance of LSAM-ML in DGM 2.

Study 4a: Absolute bias of SEM-ML for DGM 1

	Sample size						
beta	**N=50**	**N=100**	**N=250**	**N=500**	**N=1000**	**N=2500**	**N=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

Study 4a: Absolute bias of LSAM-ML for DGM 1

	Sample size						
beta	**N=50**	**N=100**	**N=250**	**N=500**	**N=1000**	**N=2500**	**N=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 suggest an even better relative performance of LSAM- over traditional SEM-ML estimation for smaller N and higher beta. Thus, we can see the negative small sample bias come into play in this study indirectly as well. Results look 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.

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 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 estimator. The evident performance differences observed between traditional SEM and SAM estimation in the study by Valentin Kriegmair are thus undermined. —— QUESTION: RMSE performance? Undercoverage? In context of constrained ML? What are the RMSE values in study 1? 22.6 vs. 26.4? Percent? –> Question if better performance is mostly in higher N or only in lower N. ——

Additionally, 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. The apparent performance advantages for traditional SEM in these conditions compensates the advantages in positive conditions. In context of empirical studies, where the valenced 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 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, turned out not to be one, upon closer examination. In consequence, we did not conduct a joint study and stopped the collaboration.

Discussion

Repeating results, shortly

Discussion of AC

- Necessity of structured approach? vs. Implicity of Discussion
- High additional load to already potentially high stakes
- Maybe, just having preliminary meetings to really find disagreements first, only then start collaboration (Starting with 2nd part, maybe add more explicit definition and collection of disagreement, in theory and especially in evidence!)
- Advantage of simulations vs. empirical papers: Simpler operationalizations, no paradigm differences, mostly numbers and preferred functions/ estimators etc.
- Another advantage: If in doubt, do both! Easier to resolve disagreements but additional increase in load and complexity of studies conducted.
- But: Generally advisable! Already more thought into: focus of what is important vs. not. More need for explanation (maybe already covered mostly by simulation protocols?), overall more rigor and transparency (as everything has to be understood by the adversary) –> all positive effects
- What about points in beginning? Increased generalizability and…?

—— Still to integrate 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.

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.

Into results:

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?

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.

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.

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.

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.

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, 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.