Adversarial Simulation

A case study

Leonard Kosanke

The git hash is: 385d8

```
library(here)
```

```
Warning: Paket 'here' wurde unter R Version 4.2.3 erstellt
```

```
here() starts at C:/Users/leona/OneDrive/Dokumente/GitHub/AdversarialSimulation
```

```
library(knitr)
```

```
Warning: Paket 'knitr' wurde unter R Version 4.2.3 erstellt
```

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. To answer this research question, we created and followed the following procedure for the adversarial collaboration process:

Outline of the Adversarial Collaboration

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. Leonard Kosanke replicated relevant parts of the study by Robitzsch (2022), and Valentin Kriegmair replicated the study by Dhaene and Rosseel (2023). The replications included the generation of individual simulation protocols, as suggested by Morris, White, and Crowther (2019). These can be accessed in our Github-repository, under releases: https://github.com/lkosanke/

AdversarialSimulation/tree/main/. In the second part, a joint study, including a joint simulation protocol, was pursued. Here, the main part of the adversarial collaboration took place. Each step of the individual simulation studies (from the substantive research question up until the interpretation of the results) was scrutinized and debated between collaborators, using adversarial collaboration techniques. Decisions were made and documented based on the most convincing argument presented, if possible. This process started based on the results of Part 1. To facilitate a structured comparison and integration of our 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, and at the same time provides a basis for systematic comparison and synthesis for the joint simulation. Additionally, 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. 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, we conducted the joint study in a structured and formalized manner. To this end, we used the tools 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 are presented in this paper. In addition, the appendix contains a separate decision log with a detailed and complete documentation of all decisions made. Here, we summarized the results of the AC-techniques used for each step of the simulation study, as well as their consequence for the decision-making process. Another aspect of structuring the collaboration, within the decision log, lied in the way we tackle decision making in the joint study. 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 we 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, we can argue and make decisions for pragmatic reasons, as for example time needed to implement another condition that could be added. 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, we might resort deciding at random. Lastly, there might be other reasons that lead to a decision, that we do not anticipate as of now or did not deem to be likely enough to get their own category for our purpose.
These four reasons grounded our decision making. In some cases, more than one of these reasons were present. In such cases, we ranked their relevance to a decision, 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

We documented and evaluated the adversarial collaboration in Part 2 (Joint study) from each collaborators subjective perspective using semi-structured diary entries based on Shah and Leeder (2016), after each step. This had two purposes: Firstly, accumulating evidence for the evaluation of the AC and the collaboration procedure we propose. Secondly, we wanted to capture information that can help us improve this procedure. Thus, the resulting diaries were used to evaluate the applicability of the adversarial collaboration and answer the research question. The evaluation of the diary entry was conducted in a semi-structured and comparative manner, resorting for example to analysis of quantity of words.

Methods for Replicating Robitzsch (2022)

The structure of this section closely aligns to our agreed upon structure of simulation studies. 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?

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. Overall, I conducted 6 simulation studies. Their 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 of 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	Yes	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. Additionally, the amount of misspecification was varied, either via different numbers of unmodelled residual correlations, cross-loadings, or both. In Studies 1b and 4a, some population values for model parameters 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 author does not give reasoning for this decision himself.

Method Selection

I included a smaller sample size N=50 for all studies, to be able to answer our substantive research questions in more detail. 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: Study 1b explicitly investigated the small sample bias of LSAM estimation in low sample sizes with N=50 and N=100 and varying lambda and phi values. Here, 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. 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). They were, however, extended to look at the interaction of beta and N for the 5-factor regression model, as this again was of interest four our substantial research questions.

I did not include a detailed mechanism to capture model convergence as detailed in the first substantive research question. As Robitzsch (2022) argue 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 estimation 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 full result analysis is available here: https://github.com/lkosanke/AdversarialSimulation/blob/main/LK/results.pdf. The same folder contains all simulation scripts that were used. The simulation was repeated 1500 times for each Study.

The first four studies investigated the comparative robustness of SAM vs. SEM estimation in a 2-factor-CFA model under different data generation mechanisms and conditions. Studies 1, 2, and 3 varied the factors sample size and different kinds and levels of misspecification for multiple estimators of either estimation method. Study 1b explicitly investigated the small sample bias of LSAM estimation in low sample sizes, by varying lambda and phi values. Studies 4 and 4a then investigated the comparative performance of the same estimators in a 5-factor model with different beta weights, as well as again under different conditions of sample size and misspecification in the data generating mechanism. 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). They were, however, extended to look at the interaction of beta and N for the 5-factor regression model. Different performance measures were computed and compared between estimators, conditions and with the original papers results. I did not include a detailed mechanism to capture model convergence as detailed in the first substantive research question. As Robitzsch (2022) argue 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 estimation 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 full result analysis is available here: https://github.com/lkosanke/AdversarialSimulation/blob/main/LK/results.pdf. The same folder contains all simulation scripts that were used. The simulation was repeated 1500 times for each Study.

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.