Project Proposal for a Subproject in the DFG funded Priority Programme META-REP (A Meta-scientific Programme to Analyse and Optimise Replicability in the Behavioural, Social, and Cognitive Sciences)

# Enhancing the Robustness of Observational Social Science Research by Computational Multi-Model Analyses


## Principal Investigators:

**Prof. Dr. Katrin Auspurg**
LMU Munich, Department of Sociology, Chair of Quantitative Methods of Empirical Social Research

**Dr. Andreas Schneck**
LMU Munich, Department of Sociology, Chair of Quantitative Methods of Empirical Social Research


## 1. Abstract


Replication sciences have so far heavily focused on experimental research where uncertainty of results is primarily caused by sampling error. However, the analysis of the robustness of research with non-experimental data, which is dominant in many social sciences disciplines, requires other methods that also capture uncertainty caused by model choice: In research with non-experimental data, there are often numerous possibilities to specify the analysed samples, the functional form of studied associations, the selection of covariates, and regression models. In addition, unobserved heterogeneity can endanger the validity of non-experimental research (so-called sensitivity). However, large-scale evaluation studies are missing, and there is also a lack of suitable methods for this purpose.

This project therefore asks: How can the robustness of non-experimental social science research be assessed and improved with the help of computational "multi-model" programs? Three closely related research objectives serve this purpose:

(1) The further development of promising programs for robustness and sensitivity analyses (so-called "multi-model", "multiverse", or "specification curve" analyses) for their use in large-scale evaluations. For example, we aim at developing standardised robustness measures and defining model variants to be tested (in the form of sample/variable/regression models).
(2) To achieve the first large-scale robustness analysis of effect estimates with regression analyses of non-experimental data. For this purpose, the tools developed in (1) will be applied to 100 studies published in leading journals of relevant disciplines (sociology, political science, and economics). We will investigate the *reproducibility* rate: To what extent are results reproducible with the models and data of the primary studies; and what role do possible (coding) errors play in this? The *robustness* rate: To what extent are results robust against (which) alternative models? As well as the *sensitivity* rate: To what extent does unobserved heterogeneity threaten the robustness and validity of estimated effects? These comprehensive analyses also allow for the first time a systematic identification key (statistical) sources for robustness.

(3) The exploration of routines to improve robustness in primary research: To what extent can multi-model and sensitivity analyses already help researchers to arrive at more robust estimates? As another novelty we will implement with "robustness notes" a new publication format.

With these three closely intertwined research objectives, the project makes important contributions to the "what" and "how" question of the META-REP Priority Programme: What is the replication rate (robustness), how can it be determined, and how can robustness be improved already in primary research?

## 2. Background, motivation, scope and main work packages of our project

Although experiments are on the rise in the social sciences, quantitative analyses in disciplines such as sociology, political sciences, and economics still mostly rely on observational data (Green & Gerber, 2003). The research goal is very often to understand how a treatment variable of interest (such as social inequality) affects a specific outcome (such as happiness). These analyses not only suffer from sampling error, but also from a large amount of "model uncertainty": There are different ways to prepare data (e.g. how to handle "outliers"); there are multiple ways to operationalise concepts such as social inequalities (e.g. focusing only on income and/or assets; using the variance, ratios, or more complex measures; c.f. O'Brien, 2018), and there are many ways to specify a regression model (e.g. regarding the selection of covariates). This multiplicity of data analysis strategies is prone to becoming a "garden of forking paths" (Gelman & Loken, 2013): Researchers can and do run many variants of models. There is often enormous ambiguity about which model to choose, and researchers may misuse their degrees of freedom to "cherry-pick" only models that lead to "significant" results (Simonsohn, Simmons, & Nelson, 2019).

There are meanwhile large scale replication audits for experimental social science studies (e.g. Open Science Collaboration, 2015), but none for observational studies (except the SCORE project we will collaborate with, for a description, see Section 2.2). Therefore, we do not know: How replicable are social science findings based on observational data? To find out, we need methods that do not only reflect uncertainty that is caused by sampling error (which is the main focus of replication audits for experiments), but also uncertainty that is caused by model choice (Freese & Peterson, 2017). For experimental studies, the silver bullet to test replicability would be to run the same analyses with a different random sample (so called "direct replication"). However, a misspecified model would work on all samples equally "well" (O'Brien, 2018). Consequently, the crucial task of testing the robustness of observational studies is to test the replicability not across different samples, but different specifications of regression models.

Therefore, we ask: *How can we analyse and enhance the robustness of observational social science research?* We will pursue this aim with three closely related objectives.

(1) *Refine computer tools for robustness and sensitivity analyses for large-scale analyses*

*Robustness* means the extent to which analyses replicate when using the same data as original papers, but different methods to prepare and analyse the data (see the terminology defined in the META-REP proposal). In recent years, several promising computational tools for robustness analyses were developed, that we refer to as "multi-model" analyses (Muñoz & Young, 2018). These estimate the distribution of treatment effects across the entire space of alternative regression models that seem reasonable (e.g.

defined by different specifications of analyses samples and covariates). This allows to determine the extent to which estimates are non-robust to different model choices. *Sensitivity* is a special variant of non-robustness to unobserved variables (Imbens, 2003): Even unobserved heterogeneity might invalidate estimates. There are promising computer tools for sensitivity analyses (Harada, 2013) as well. However, all these tools need refinements before they can be used in large-scale assessments. We try to achieve this by e.g. developing standardised robustness measures and defining reasonable model spaces that should be tested.

(2) *Apply these tools to provide the first large-scale assessment*

We will apply the tools developed in (1) to 100 social science estimates based on observational data with regression analyses to achieve the first large-scale assessment of their reproducibility, robustness, and sensitivity rates. We will start with the *reproducibility*, i.e. to what extent remain results stable when re-running syntax files on the data used in original research, with and without fixing possible errors in the original syntax files and data. After that, we estimate the *robustness* and *sensitivity* rates and identify risk factors for non-robust and sensitive results. We will sample for these analyses core estimates published in leading international journals in three disciplines that frequently work with observational data: sociology, political sciences, and economics.

(3) *Develop recommendations for enhancing the robustness of original estimates*

With few exceptions, computational robustness and sensitivity analyses have so far only been used in meta-research. Ideally, however, they would already improve original research. This could possibly be achieved by implementing them by default in publication and review processes. We will survey authors and further experts to see whether review-processes informed by robustness and sensitivity analyses could help authors to arrive at more restricted model choices and hence more robust, insensitive results. As another practical tool to improve (the transparency on) robustness, we will implement a new publication format ("robustness reports").

With these three closely intertwined research objectives, we contribute to the "what "and "how" question of the META-REP priority programme: What is the replication (robustness) rate, how can we find out, and how may we best improve replication rates in future. All our objectives are pursued with standardised tools that economise on research resources and provide a more systematic overview on the robustness of observational social science findings than it was possible with the previous literature.

**For more information**, please contact:

Katrin Auspurg (Katrin.auspurg@lmu.de

Andreas Schneck (Andreas.schneck@lmu.de)