Proposal for Research and Development of Research Tools:

“The Role of Theory in Understanding and Resolving the Reliability Crisis”

Nate Breznau, breznau.nate@gmail.com

University of Bremen

SOCIUM Research Center on Inequality and Social Policy

Abstract

This is a copy of a successful German Science Foundation application proposing to conduct research on the reliability of science, and to develop applications that can help address this reliability. It focuses on the role of theory and causal inference from a meta-perspective.

Breznau, Nate. University of Bremen, Germany

The Role of Theory in Understanding and Resolving the Reliability Crisis

1. Project Description

1.1 State of the art and preliminary work

Overview

The reliability of a given empirical test is a product of the quality of the theory used to design that test. Too many unknowns or competing causal pathways from test variable X to outcome variable Y renders the results unreliable. If hypothesis tests are repeated under theoretical ambiguity, an entire area of study may be unreliable. Such a scenario could explain a lack of consensus and failures to reproduce findings in that area. In the behavioral and social sciences especially, theoretical ambiguity is acute due to the complexities of human interaction and societal organization.

Diagnosing theoretical ambiguity in a single hypothesis test involves reviewing theory and applying logical reasoning to identify sources of uncertainty and potential bias. It requires methodological rigor as well, but methods cannot fix theoretical ambiguity. Theoretical inquiry to improve a single hypothesis test is time consuming, and alone will not improve reliability across an entire area of study. Therefore, I propose an approach to maximize theory while minimizing the time investments of scholars working in an area.

The goals of this proposed project are:

- (1) Test the extent of theoretical ambiguity in one area of study, as a cause of empirical unreliability in that area.
- (2) Test if computer-assisted comparison of causal models can efficiently identify theoretical ambiguities in that area.
- (3) Test if crowdsourcing theoretical claims combined with computer-assisted causal model comparison can improve reliability and the replicability of findings in that area.
- (4) Develop computer systems to improve and economize the process for deployment across all areas of cognitive, behavioral and social sciences.

The first two goals target the question of “**why** replication rates vary” posed in the META-REP SPP to which this application is addressed. The third and fourth target the question “**how** to increase replication rates”¹. As I will discuss herein, replication and reliability are closely connected in the scientific process, therefore by addressing theoretical reliability I expect to improve empirical replicability.

Motivation

The construction of social science theory demands explanations of phenomena that transcend time and space-specific observations. The goal is to establish general truths within

¹ https://www.dfg.de/foerderung/info_wissenschaft/ausschreibungen/info_wissenschaft_20_38/index.html

certain scope conditions (like common features of neighborhoods, voluntary civic organizations or an entire society) (King, Keohane, and Verba 1994; Stinchcombe 1999). The development of “truths” in theory come from testing and re-testing to ensure “we are not dealing with a mere isolated ‘coincidence’” (Popper 1935:45). Testing and re-testing is a process of replication by definition (Christensen, Freese, and Miguel 2019). Designing replications requires theory. If theory is ambiguous, then design is lacking and results are uninformative. If replication results are ambiguous, then the theoretical “truths” derived from these results will also be ambiguous, in the sense that they are based on unreliable findings. If scholars are not aware that their theory is ambiguous, a vicious cycle takes place where faulty findings and ambiguous theory continually reproduce one-another. The outcome of this process in a given area of science is unreliability.

Theoretical ambiguity poses a unique risk to the reliability of findings in the social sciences because there are no laws, unlike the physical sciences, and because of a greater inter-subjectivity of researchers (Lucas, Morrell, and Posard 2013). This claim is nothing new. Weber and Merton wrote about types of subjectivity threatening credibility of social research long ago (Merton 1973; Weber 2002[1919]). What is “new” is a recent watershed of studies suggesting science is empirically unreliable, suggesting that Merton and Weber’s warnings were justified. Anti-scientific practices such as the obsession with p-values, publication bias, HARKing² and a general reliance on running regressions instead of focused and logical theorizing is undermining the credibility of science both in quantifiable terms and in the perception of the public and policymakers (Funk et al. 2019; Munafò et al. 2017; Pickett 2020; Young 2015, 2018). The pressure of this apparent ‘crisis’ has produced new attention to replication (Freese and Peterson 2017; King 1995; Schwarber 2019). Careful scrutiny of previous original studies through replication reduces bias, resolves open questions and identifies cases of fraud – evidence of this is already strong in all areas of psychology and gaining traction in political science, economics and sociology (Brezna 2020a).

Four broad categories can roughly define replications (Christensen et al. 2019):

Type	Data	Models
1. Verification	same data,	same test models
2. Direct Replication	new data,	same test models
3. Reanalysis	same data,	new test models
4. Extension	new data,	new test models.

In categories 2-4, an original study is altered in some way; whether through data, methods or both. Almost no study qualifies as ‘original’ in social science. All research contains elements of previous research or previous knowledge used to formulate a new hypothesis or design a new test of an old hypothesis. This means that *most existing research is a replication by definition*, even though many scholars do not label their studies a “replication”.

This phenomenon can be observed in the literature on immigration and redistributive preferences, an area I am familiar with. For example, Sides and Citrin (2007) published a study linking immigration to public preferences for redistribution. Five years later Emmenegger and Klemmenson (2013) published what would be an “extension” by definition in the list above, although they did not call it a “replication”. They theorized that motivation is a key factor moderating the role of immigration in shaping preferences for redistribution, so they added motivation

² Hypothesizing after results are known.

measures to the original model formulated by Sides and Citrin and re-ran the analyses. I would argue that most of social science works this way and could be labelled “replication”. Therefore, an enquiry into the what, why and how of replications (the goal of the SPP), is an enquiry into science at a basic level.

Categories 2-4 in the types of replication *all require some amount of theory*, even if just a basic suspicion about why a previous study was wrong, unreliable or should be repeated. This brings me to the idea that investigating theoretical ambiguity in already existing studies will address “why” replications are or are not successful, or “why” social science appears to be in a replication crisis. If preexisting studies contain theoretical ambiguity then future studies based on them are more likely to be unreliable, regardless of whether these studies are labeled “replications”. Moreover, if by definition many studies in an area already are replications by definition, then it is time to investigate them for what they are rather than search only for studies that refer to themselves as “replications”. Moreover, when combined with efforts of others to conduct new replications, looking at pre-existing studies in comparison should help triangulate reproducibility crisis problems.

In the Emmenegger and Klemmenson (2013) replication, they added variables to the original model, thus making theoretical improvements to it. In doing so, they made other theoretical assumptions *that everything else about the original model was correct*. Although most scholars make this assumption implicitly by simply copying a model from before and changing one aspect, Emmenegger and Klemmenson did so explicitly, stating that they followed Sides and Citrin’s work, “in order to safeguard the analysis from allegations of model misspecification and omitted variable bias” (2013:234).

Their statement is an example of what I understand to be the status quo. That scholars are unlikely to consider every ‘other’ independent variable in a model when conducting their replications (again whether they call them replications or not). As in this example, scholars tend to shift the theoretical burden away from themselves and on to the previous study. If some initial previous study was theoretically misspecified in some way, the reproduction of this misspecification is unreliable and any studies based on it thereafter. As I discuss in the next section, this type of practice of simply accepting other studies’ causal implications as correct, could be a key explanation for a lack of reproducibility or consensus across the social sciences.

State of the Art

Formal logic using special mathematical notation provides a system for reducing theoretical ambiguity by directly comparing the claims of statements (a system at least as old as Aristotle, and what some refer to as “mathematical social science”). It has strict rules for determining if a given model can test the causal relationship specified in a hypothesis, in other words if the test is “identified” logically (Pearl 2010). If forced to consider all logical implications of a model, social scientists would not automatically accept previous models as theoretically correct. By scrutinizing all implicit and explicit theoretical arguments with formal logic they would reduce the instance of non-interpretable and unidentified results (Fredericks and Miller 1985)³. The problem is that the use of formal logic in the social sciences is extremely rare, except for economics. This is unlikely

³ I intentionally do not engage in the philosophy of logic here. It is debatable and has many schools, and this debate spills over into social science (Collins 1989); but as the use of formal logic is rare, I skip this debate and focus on causal inference as an efficient, albeit certainly not perfect, way forward to address the current reliability crisis.

to change, as I doubt any individual or agency can force social scientists to implement formal logic on a mass scale.

There is; however, a rapidly growing alternative. Causal inference provides an extension of formal logic where researchers can draw their theories, and with the help of a special type of calculus, often implemented with a computer, they can quickly deduce the model's implicit logical and, thus, mathematical assumptions. This means they can quite easily determine if the test can identify a causal effect (Elwert 2013; Pearl 2010). In fact, this graphical causal inference approach is similar to methods of drawing structural causal models developed over 100 years ago and applied heavily in psychometrics, education research and the study of social mobility (Bollen and Pearl 2013; Wolfle 2003; Wright 1934). Thus, in speaking about “causality” herein I refer to both causal inference and structural causal models, as they have causal identification in common (Bollen and Pearl 2013).

Causal inference reveals if a test can identify a causal effect of X on Y . This method puts the burden on theoretical claims. Each arrow and each absence of an arrow in a causal diagram *must* have theoretical defensibility or the model is untrustworthy and presumed to be wrong until theoretical arguments are given (Keele, Stevenson, and Elwert 2020). The practice of throwing ‘control’ variables haphazardly into regression models does not follow a logic of causal inference, yet it is a common practice (Carlson and Wu 2011). In the Emenegger and Klemenson example, they simply accepted that the previous study is correct except for one variable without further theorizing (the previous study had eight other variables). Of course, many studies do not present their models as “causal”, but this should already indicate why they cannot be replicated. If scholars have no idea what causes the estimated statistical effects in their studies, they should not be surprised when others cannot replicate those effects. Thus, if social science is to take the replicability crisis seriously, it is time to take on causality directly.

Methods of meta-analysis, in particular comparing model specifications and testing how they impact results are emerging across the social sciences and can reduce some causal uncertainty (Bresnau 2019a; Simonsohn, Simmons, and Nelson 2015). Researchers can also simulate results from different causal assumptions to help identify potential conflicts or assess the robustness of models under causal uncertainty (Cox 2016). Although useful, these methods compare causal models on their **outputs**. I suggest comparison on the **input** side instead. *If the underlying causal models are not correctly identified, then output results are unreliable, no matter how many are meta-analyzed or simulated.* Social science cannot resolve which model is correct by running nine billion more regressions, because most of those regressions are likely to contain causally impossible theories (Muñoz and Young 2018).

In Figure 1, I offer an example from my own research area on linkages between immigration and public preferences for income redistribution to illustrate the possibility of identifying unreliability by comparing causal models. The box in the lower right corner lists regressions used in three hypothetical studies. The panels for each study (A, B and C) represent a causal model given those regressions and some basic logic.

Study A & B in Figure 1 are identical except for national economic productivity in Study B's regression. Productivity is a potential cause of immigration as indicated by the “?” in the causal model, given that stronger economies attract immigrants. If true, then Study A's regression results cannot be trusted, and the causal model for Study A must be incorrect because economic productivity causes omitted confounder bias (Rohrer 2018). In other words, a statistical test might

suggest a ‘causal’ effect of immigration in Study A, but it could be an artifact of their shared co-cause and is not reliable as such.

Study C has immigrant integration policies instead of productivity as an additional variable. There is good reason to suspect that the dependent variable public preferences and the independent variable number of immigrants *both cause* these policies. Immigrants create new social dynamics and political constituencies calling for new policies, and self- and group-interests may be triggered by immigration leading to a desire to safeguard native resources by reducing redistributive policies, also known as welfare chauvinism (Breznau and Eger 2016). If true, having the integration measure in the regression introduces a collider bias and could remove the causal effect being investigated through suppression. In the literature on immigration and policy preferences, economic productivity and integrative policy appear in many regression models (Brady and Finnigan 2014; Eger and Breznau 2017). Looking at the three causal diagrams in Figure 1 means that resolving whether an arrow exists where the question marks are located (Study B & C) is absolutely necessary before putting any faith in results in this area of research.

Figure 1. Comparing Causal Path Models to Identify Theoretical Ambiguities

Figure 1 is a major simplification, to illustrate my point. In reality, the literature has highly complex models that would require large causal diagrams. Therefore, using machine guided algorithms to draw and compare them is necessary, but promising as an efficient way forward. If a

⁴ Constant and any other independent variables not presented for ease of interpretation.

computer were given the causal models in Figure 1, it would determine that if all arrows in all models are theoretically causal, then Model A and Model C must be wrong due to omitted confounding of economic productivity. However, if the question mark path in Model B is not causal, then only Model C would be determined to be wrong. There are far more causal considerations beyond colliding and confounding and Judea Pearl offers a system of “do-calculus” to treat all the different forms of identifying causality (Pearl 2009); again this is a simplified example.

The computer cannot decide if a path is causal or not. This requires human theory, and theory construction often takes months or years. Thus, I propose to test and develop a method of crowdsourcing theory requiring small time investments from many scholars working in the area – many already in my personal network, or having already participated in crowdsourced work with me (see Own Preliminary Work section). Again, a primary motivation here is to implement solutions that are realistic, rather than hope that all social scientists suddenly become causal logicians and spend years developing theories that are complementary. Assuming successful development of a method to engage scholars in a way that they need not necessarily invest heavily in new methods or epistemologies (see Objectives section), their theoretical arguments could be fed back to the computer to improve its comparisons and conclusions about causal assumptions using the logical mathematics of causal inference.

Over time this would build a robust indicator of which studies’ assumptions are more reliable than others and which studies are likely wrong and should be discounted – for example removed from meta-analyses or excluded from future replication efforts. Ideally, it could even quantify the role of theory in the reliability crisis because results of studies and their causal assumptions could be meta-analyzed by a computer to see if models that have identified causal effects (or at least no known threats to their reliability) produce more consistent results.

To pursue this concept, I propose a project focusing at first on a particular hypothesis that I know best: that immigration reduces support for income and other forms of material redistribution. This area is already known to be rather unreliable and has easily accessible public data sources across most studies (as I discuss next). If initial tests and computer system development goes well, I would plan to deploy the method across cognitive, behavioral and social science disciplines. But first, a test in my own area.

Own Preliminary Work

I have extensive experience studying immigration and social policy preferences – for example, preferences for more or less income redistribution, labor market protections and restrictive immigration across many countries of the world. It is known to be an unreliable area because results point in all different directions. Of those studies with ‘significant’ effects, sizes go from roughly zero up to more than a half of a standardized effect (very large) *in both directions* and with varying levels of significance from barely $p < 0.05$ to $p < 0.001$ (Auspurg, Brüderl, and Wöhler 2020; Brady and Finnigan 2014; Breznau and Eger 2016; Burgoon 2014; Eger 2010; Hjerm and Schnabel 2012; Schmidt-Catran and Spies 2016). A naïve policymaker reviewing all statistical test results in this area would conclude that they are of little use to her decision-making process. She might even question why we fund such research in the first place. In my view, the logic of such an argument is valid, and it constitutes a reliability problem, if not crisis.

This project I propose draws on work I developed in 2018, in a study called ‘The Crowdsourced Replication Initiative’ (“CRI”) where collaborators and myself got together 153

researchers in 71 teams, who all used the same data to test the same hypothesis that immigration reduces support for redistributive social policies (Brezna, Rinke, and Wuttke 2019). The rate of a positive effect of immigration on social policy within a 95% confidence interval was 28%, while the rate of a negative effect was 17%. The remaining 55% were either below 95% confidence or in a few cases, the tests failed. The results from the CRI provide over 1,200 plausible data-generating models and their test outcomes. “Plausible” means the participating researchers in the CRI submitted what *they believed to be the most correct model or set of models*. If funded, my project team can immediately start converting these statistical models into causal path models and immediately develop machine-based comparisons, as these models reflect what one might expect to find in the published literature. Therefore, the proposed project already has data to work with.

I am a social scientist by training but I am now familiar with the basics of machine learning, deploying web-based applications (with Java, Python, docker and a server) and text mining. In the CRI, we tried to explain the variance in outcomes across the teams using logic and some exploratory analyses, but to check that we did not miss anything we implemented machine algorithms to determine which model specifications led to which results. In the CRI, I developed interactive solutions for presenting massive amounts of data to users, in a shinyapp example I created with collaboration from my current PhD student, the user can interact with different model specifications ([link to shinyapp](#)). My current DFG funding⁵ led to the development of what I call the *bibfilter*, a web-based, searchable bibliography collecting a list of all studies of attitudes toward redistribution. It provides a starting point for investigating the literature with systematic review, with up to 100 studies on the topic ([link to bibfilter](#)). Finally, in my current DFG project we are implementing machine driven text mining and sentiment analysis of the German parliamentary activities to link their contents to public opinion. This text-mining method will be necessary to train a machine to automate some of the process of collecting data in articles in the proposed project. I submit this review of my own work as evidence that I am fit to lead the development of such applications with the help of experts. Moreover, it reflects a new approach to doing science; I believe the future of social science necessitates some degree of data and computer science applications – namely, in order to address the reliability crisis.

Beyond my methodological and computer-based knowledge, my work on the replication crisis, based on an analysis of the CRI data reveals that 3 replicators per study would be necessary to really achieve reliable replications⁶. This means every scholar would have to replicate three other studies per study she publishes to have replication reliability in an area; an impossible expectation of scholars. Therefore, economizing this process is absolutely necessary. As the principal investigator of the CRI, I am also aware of the challenges associated with coordinating hundreds of academics. But aware that it is possible, and that they would give small amounts of time to address reliability in their own areas. Moreover, my recent publications all reflect a major shift in my scientific career toward open science, dealing with credibility crises in science and trying to do science as more of a communal effort. I recently started teaching an open science course and have given two invited talks on the topic of open science (see my attached CV and [blog](#)). This reflects my passion for the topic of this grant.

⁵ Project number 401013559, <https://gepris.dfg.de/gepris/projekt/401013559>

⁶ “How Many Replicators Does It Take to Achieve Reliability? Investigating Researcher Variability in a Crowdsourced Replication Study.” Under review: <https://osf.io/zvepn/>

Finally, I want to stress my capacity to use grant funding successfully. After 1 year and two months of my current DFG project, I produced two solo-authored publications (Breznau 2019c, 2020d) and one under review (Breznau 2020c) from the funding. Given the Covid-19 pandemic, my team shifted focus and now have an in-press article on government pandemic policy interventions and public preferences (Nguyen, Breznau, and Heukamp 2021), and another in the review process on income inequality, government response and public perceptions (Breznau, Heukamp, and Nguyen 2020). Given that I also learned qualitative text analysis, machine learning and basic web applications in the same period, and published articles (see CV) and data⁷ not directly related to that project, I would consider this successful usage of public science funding.

1.2 Project-related publications

1.2.1 Articles published by outlets with scientific quality assurance, book publications, and works accepted for publication but not yet published.

1. Breznau, Nate. 2020a. "Does Sociology Need Open Science?" *Societies*. In-press. Access: <https://osf.io/preprints/socarxiv/d37be/>
2. Breznau, Nate. 2020b. "I Saw You in the Crowd: Credibility, Reproducibility and Meta-Utility." *PS: Political Science and Politics*. In-press. Access: <https://osf.io/preprints/socarxiv/aex5z/>
3. Breznau, Nate. 2016. "Secondary Observer Effects: Idiosyncratic Errors in Small-N Secondary Data Analysis." *International Journal of Social Research Methodology* 19(3):301–18. Access: <https://osf.io/hzm9n/>
4. Breznau, Nate. 2015. "The Missing Main Effect of Welfare State Regimes: A Replication of 'Social Policy Responsiveness in Developed Democracies' by Brooks and Manza." *Sociological Science* 2:420–41. Access: <https://www.sociologicalscience.com/articles-v2-20-420/>

1.2.2 Other publications, both peer-reviewed and non-peer-reviewed

5. Breznau, Nate, Eike Mark Rinke, and Alexander Wuttke et al. 2019. "Crowdsourced Replication Initiative: Executive Report." University of Mannheim. Access: <https://osf.io/preprints/socarxiv/6j9qb/>
6. Breznau, Nate, and Maureen A. Eger. 2016. "Immigrant Presence, Group Boundaries, and Support for the Welfare State in Western European Societies." *Acta Sociologica* 59(3):195–214. Access: <https://osf.io/qe6yk/>
7. Eger, Maureen A., and Nate Breznau. 2017. "Immigration and the Welfare State: A Cross-Regional Analysis of European Welfare Attitudes." *International Journal of Comparative Sociology* 58(5):440–463. Access: [download link](#).

1.2.3 Patents

1.2.3.1 **Pending:** not applicable

1.2.3.2 **Issued:** not applicable

2 Objectives and work programme

2.1 Anticipated total duration of the project

Initial project anticipated at 36 months. With a successful study and prototyping computer systems to deploy across disciplines, a second 36 months would be sought.

⁷ <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/IVKYIE>

2.2 Objectives

To form a research team of myself, a postdoc who specializes in causal inference and has theoretical and empirical knowledge relevant to immigration and policy preferences, and a doctoral student who will study a social science degree but focus on computer and data science methods.

I would aim to achieve the following objectives (see Table 1 below for a time plan):

To **identify all studies** investigating the role of immigration in the formation and change in preferences for redistribution and redistributive social policy. Certain parameters would apply, for example no predatory journals, possibly no studies published in Elsevier journals (out of solidarity in case German universities continue their boycott via *ProjektDEAL*). As stated in the Preliminary Work section, the *bibfilter* provides a list that will make this a much easier task.

To **create causal diagrams** from identified existing studies. These will come **from the CRI** (described in Preliminary Work) and **from identified published studies** using the *bibfilter* and systematic review. This will involve development of *R Studio* and *Stata* packages that link regression output with dagitty applications (Textor et al. 2016; Wu 2019). This can be achieved with work from the PhD student and a Master student assistant. This would be an early contribution (within the first 6-12 months) of the project by deploying these packages for public use, and quickly provide the project with a large sample of causal diagrams on which to test the proposed methods. For the other identified studies, causal diagrams will be encoded by hand from the regression models. For studies with relatively basic regression models, we should be able supervise a Master student to help with this task, and ideally train a computer to make it easier and more accurate. However, for studies that are more complicated (multilevel models, structural equation models, qualitative comparative analysis and potentially qualitative research⁸) close supervision, if not self-coding will be necessary on the part of the postdoc and myself. We will carefully monitor the amount of time needed to add each study to estimate what we can realistically achieve in three years, and constantly cross-check each other's work for errors.

To **develop a computer system** for storing and comparing causal diagrams. The PhD student will take a leading role with the help of research assistants from computer science. The comparison will require **methodological development** of a quasi-new method, and this part demands the postdoc and my own causal inference skills and experience in the field ensure that the comparisons work, and a skilled PhD student to make sure the computer system works. The method is simple at its core (see Figure 1) and uses the logic of causal inference in one causal model, but applies it to two or more causal models to determine where their assumptions are in direct conflict. This is made possible because computer programs such as dagitty output assumptions in the form of standard mathematical language. For example, if one model has implication " $A \perp D \mid E$ ", indicating that A and D are conditionally independent given "E", and another model has " $A \rightarrow D$ ", indicating A is a cause of D, then the computer could flag these two models as referring to incompatible realities (see dagitty.net to see how single causal models are analyzed).

Two primary outcomes will follow (a) whether the test parameter (in this case immigration) can be considered causal, and (b) where models cannot co-exist causally (as demonstrated in Figure 1). This means extending the **methodological development** to determine exactly how the

⁸ Research using qualitative methods could be included in this planned project, but it will be excluded at first for practical reasons. Certainly qualitative research involves theory and hypothesis testing and therefore causal models can be extracted, but it lacks the power of counterfactual statistics, thus test results and replicability are extremely hard to compare across studies. This would be a feature if the project goes on to a second stage of funding.

computer can enhance theory; for example, questions that can only be solved through testing such as: how many studies are necessary to generate useful output? How much of the causal diagramming can be computerized and how much still relies on human causal diagramming? How much knowledge of the field and of causal inference is necessary to do the causal diagramming? How much knowledge of causal inference is necessary for users to implement the results of our new tools (what are the costs to the user of this proposed method)?

To **train the computer system**, to output conflicts between causal models and to list assumptions that are common across causal models without error. The conflicts are obviously something that needs resolution before seeking to place faith in any particular findings from the field; however, the common assumptions that are not in conflict should also have supporting theoretical logic. Therefore, the postdoc and myself will review these assumptions as they are produced in the project and determine if they are sound assumptions or may need review by experts, or even further specialized testing.

To **crowdsource experts** in the field to review and comment on theoretical assumptions deemed important through our computer algorithm. This requires much testing and will involve all project members first as testers themselves and then local experts the University of Bremen where we can do cognitive pretesting of computer-assisted surveys, deliberations or possibly choice experiments to learn how to gain maximal theoretical information with minimal time and resource commitment from experts outside of our professional contacts. The University of Bremen is an ideal location to carry this out in-person because of the extensive research on redistribution, immigration and especially redistributive social policies.

To **quantify the role of theory** in the variation (lack of consensus) and reliability (rate of models that are misspecified and cannot test the hypothesis) of results in this test case. At first this project will provide detailed information about **why and how the reliability crisis impacts one area of political and social science**. Then to estimate how much of this role of theory can be improved through a computer-assisted meta-causal tool and at what cost. In particular, to **test if it is deployable** across all disciplines for example those using more complicated models (for example with systems or simultaneous feedback), experimental studies or massive data sources such as human genomes or neural mapping.

2.3 Work programme including proposed research methods

Table 1: Analytical Steps Overview

Months	PI	Postdoc	Doctoral student
1-6	<ul style="list-style-type: none"> • Refine causal inference skills • Create causal diagrams from CRI • Supervise project, integrate doctoral student into PhD studies • Hire MA and BA researchers • Supervise development of R Studio and Stata packages to 'harvest' causal models from existing studies 	<ul style="list-style-type: none"> • Refine causal inference skills • Create causal diagrams from the CRI studies • Begin systematic review • Investigate theory in the immigration and preferences literature • Hire MA and BA researchers 	<ul style="list-style-type: none"> • Integrate into BIGSSS program • Experiment with computer system development • Computer science training • Begin development of R Studio and Stata packages to 'harvest' causal models from existing studies (use dagitty.net)
6-12	<ul style="list-style-type: none"> • In-depth comparison of causal models, produce rules for a computer to follow • Develop method as a whole; in depth documentation; public platform for sharing (GitHub); write a methods paper 	<ul style="list-style-type: none"> • In-depth comparison of causal models, produce rules for a computer to follow • Develop method; in depth documentation; public platform for sharing (GitHub); methods paper • Investigate theory in immigration and preferences literature • Finish systematic review 	<ul style="list-style-type: none"> • Implement rules and decision-criteria in new computer platform • Finalize development of R Studio and Stata packages to 'harvest' causal models • Begin training computer system • Investigate machine learning to extract causal information from articles

13-18	<ul style="list-style-type: none"> • Finish CRI studies causal comparisons; publish results 	<ul style="list-style-type: none"> • Finish CRI studies causal comparisons; publish results • Begin comparing identified studies • Begin cross-validating research with other SPP projects 	<ul style="list-style-type: none"> • Finish CRI studies causal comparisons; publish results • Work with research assistants to create infrastructure to deploy across disciplines • Develop dissertation project
19-24	<ul style="list-style-type: none"> • Finish identified studies comparison • Decision fork <ul style="list-style-type: none"> (A) The project looks relevant for deploying across disciplines, start applying for second phase funding (B) Project is not transferrable, make plan to publish area-specific results and any methods findings 	<ul style="list-style-type: none"> • Finish first round of identified studies comparison • Compare CRI studies and identified studies; meta-analysis; publish results • If (A) learn grant writing skills, contribute to grant • If (B) focus on area-specific results and publications • Offer workshops on theory and causal inference 	<ul style="list-style-type: none"> • Work with research assistants to create infrastructure to deploy across disciplines • If (A) learn grant writing skills, contribute to grant and determine all resources necessary to deploy computer system across disciplines • If (B) finalize area-specific technologies to improve area-specific future research
25-30	<ul style="list-style-type: none"> • If (A) finish grant proposal • Feasibility check – how may more identified studies can we encode and compare? 	<ul style="list-style-type: none"> • Continue with identified studies comparison • Work on publications 	<ul style="list-style-type: none"> • Focus on dissertation project • Continue supervision of computer system deployment either locally (B) or globally (A).
31-36	<ul style="list-style-type: none"> • Draw conclusions on role of theory • Finalize comparison of as many studies as possible • (After 36 months) continue working on the role of theory in replicability 	<ul style="list-style-type: none"> • Draw conclusions on role of theory • Depending on interests publish what social science can gain from causal inference • Generate computer output: where is attention to theory needed most 	<ul style="list-style-type: none"> • Finish dissertation project • Contribute to all project publications • Generate computer output: where is attention to theory needed most

2.4 Collaboration within the priority program

If funded I would aim to collaborate with the work proposed by Alexander Röhm, we could both be external validity checks for our approaches to meta-comparing theory in existing studies. This is important because his area of research, like most of the projects in the SPP, is very different from my own. This provides the opportunity to pursue my goal of having a method that works regardless of the subject.

There are several projects from the SPP proposing a ‘many analysts’ or ‘many labs’ approach and any/all of these projects could provide potential testing grounds for deploying my proposed system, especially after the first year, or if a second phase of funding were acquired. These projects will produce large amounts of standardized data, methods and results.

Crucially, I see the principal investigators of the META-REP SPP as potential collaborators because they plan to develop extensive infrastructure to promote their goals. This infrastructure could be essential to my proposed project. Of course, I am likely to develop it in Bremen at first, but when taking the next step to deploy it across diverse fields I would expect to make it part of their computer systems infrastructure to benefit the META-REP SPP as a whole. This means the funded researchers on my project could be highly useful to the development of their infrastructure – I could imagine even starting this collaboration right away to be maximally beneficial to our mutual goals if the SPP principal investigators thought this was a good idea.

3 Bibliography concerning the state of the art, the research objectives, and the work programme

- Auspurg, Katrin, Josef Brüderl, and Thomas Wöhler. 2020. "Does Immigration Reduce the Support for Welfare Spending? A Cautionary Tale on Spatial Panel Data Analysis." *American Sociological Review*.
- Bollen, Kenneth A., and Judea Pearl. 2013. "Eight Myths about Causality and Structural Equation Models." Pp. 301–328 in *Handbook of Causal Analysis for Social Research*, edited by S. L. Morgan. Dordrecht, Netherlands: Springer Science & Business Media.
- Brady, David, and Ryan Finnigan. 2014. "Does Immigration Undermine Public Support for Social Policy?" *American Sociological Review* 79(1):17–42.
- Breznau, Nate. 2019a. "Behind the Specification Curve." ASA Poverty, Inequality and Mobility Newsletter. Link: <https://crowdid.hypotheses.org/220>
- Breznau, Nate. 2019c. "The Underlying Public Attitude toward Government Responsibility to Intervene in Socio-Economics, Thirty-Years of Evidence from the ISSP." *International Journal of Sociology* 49(3):182–203.
- Breznau, Nate. 2020c. "Bi-Directional Feedback to the Welfare State and Public Opinion." Under Review. Access: <https://osf.io/preprints/socarxiv/h8aqe/>
- Breznau, Nate. 2020d. "The Welfare State and Risk Perceptions: The Novel Coronavirus Pandemic and Public Concern in 70 Countries." *European Societies* 0(0):1–14.
- Breznau, Nate, Lisa Heukamp, and Hung Nguyen. 2020. *The Swedish Paradox Explained? Investigating the Role of Economic Inequality and Risk Perceptions in the Novel Coronavirus Pandemic*. SocArXiv. Access: <https://osf.io/preprints/socarxiv/re7sn/>
- Burgoon, Brian. 2014. "Immigration, Integration, and Support for Redistribution in Europe." *World Politics* 66(3):365–405.
- Carlson, Kevin D., and Jinpei Wu. 2011. "The Illusion of Statistical Control: Control Variable Practice in Management Research." *Organizational Research Methods*. doi: 10.1177/1094428111428817.
- Christensen, Garret, Jeremy Freese, and Edward Miguel. 2019. *Transparent and Reproducible Social Science Research*. Los Angeles, Calif.: University of California Press.
- Collins, R. 1989. "Sociology - Proscience or Antiscience." *American Sociological Review* 54(18109):124.
- Cox, Louis Anthony. 2016. "Quantifying and Reducing Uncertainty about Causality in Improving Public Health and Safety." Pp. 1–64 in *Handbook of Uncertainty Quantification*, edited by R. Ghanem, D. Higdon, and H. Owhadi. Cham: Springer International Publishing.
- Eger, Maureen A. 2010. "Even in Sweden: The Effect of Immigration on Support for Welfare State Spending." *European Sociological Review* 26(2):203–217.
- Eger, Maureen A., and Nate Breznau. 2017. "Immigration and the Welfare State: A Cross-Regional Analysis of European Welfare Attitudes." *International Journal of Comparative Sociology* 58(5):440.
- Elwert, Felix. 2013. "Graphical Causal Models." Pp. 245–272 in *Handbook of Causal Analysis for Social Research*, edited by S. L. Morgan. Dordrecht, The Netherlands: Springer Science & Business.
- Emmenegger, Patrick, and Robert Klemmensen. 2013. "What Motivates You? The Relationship between Preferences for Redistribution and Attitudes toward Immigration." *Comparative Politics* 45(2):227.
- Fredericks, Marcel, and Steven I. Miller. 1985. "Logic and Sociological Arguments: Applications to Undergraduate Teaching." *Teaching Sociology* 12(2):149–75. doi: 10.2307/1318325.
- Freese, Jeremy, and David Peterson. 2017. "Replication in Social Science." *Annual Review of Sociology* 43(1):147–165. doi: 10.1146/annurev-soc-060116-053450.
- Freese, Jeremy, and David Peterson. 2018. "The Emergence of Statistical Objectivity: Changing Ideas of Epistemic Vice and Virtue in Science." *Sociological Theory* 36(3):289–313.
- Funk, Cary, Meg Hefferson, Brian Kennedy, and Courtney Johnson. 2019. "Trust and Mistrust in Americans' Views of Scientific Experts." *Pew Research Center Science & Society*. ([link](#)).

- Hjerm, Mikael, and Annette Schnabel. 2012. "How Much Heterogeneity Can the Welfare State Endure? The Influence of Heterogeneity on Attitudes to the Welfare State." *Nations and Nationalism* 18(2):346–69.
- Keele, Luke, Randolph T. Stevenson, and Felix Elwert. 2020. "The Causal Interpretation of Estimated Associations in Regression Models." *Political Science Research and Methods* 8(1):1–13.
- King, Gary. 1995. "Replication, Replication." *PS: Political Science & Politics* 28(3):444–52.
- King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.
- Lucas, Jeffrey W., Kevin Morrell, and Marek Posard. 2013. "Considerations on the 'Replication Problem' in Sociology." *The American Sociologist* 44(2):217–32.
- Merton, Robert K. 1973. *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago press.
- Munafò, Marcus R., Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. 2017. "A Manifesto for Reproducible Science." *Nature Human Behaviour* 1(1):21.
- Muñoz, John, and Cristobal Young. 2018. "We Ran 9 Billion Regressions: Eliminating False Positives through Computational Model Robustness." *Sociological Methodology* 48(1):1–33.
- Nguyen, Hung H. V., Nate Breznau, and Lisa Heukamp. 2021. "Locked Down or Locked In? Institutionalized Public Preferences and Government Pandemic Response in 32 Countries." *Social Policy Review*.
- Pearl, Judea. 2009. *Causality: Models, Reasoning and Inference*. 2nd ed. New York, NY, USA: Cambridge University Press.
- Pearl, Judea. 2010. "The Foundations of Causal Inference." *Sociological Methodology* 40(1):75–149.
- Pickett, Justin T. 2020. "The Stewart Retractions: A Quantitative and Qualitative Analysis." *Econ Journal Watch* 17(1):152.
- Popper, Karl. 1935. *The Logic of Scientific Discovery*. 2005 e-book. London: Routledge.
- Rohrer, Julia M. 2018. "Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data." *Advances in Methods and Practices in Psychological Science* 1(1):27–42.
- Schmidt-Catran, Alexander W., and Dennis C. Spies. 2016. "Immigration and Welfare Support in Germany." *American Sociological Review* 81(2):1–20.
- Schwarber, Adria. 2019. "Survey Shows Americans' Trust in Scientists Growing, Though Political Splits Persist." *American Institute of Physics*, August 8.
- Sides, John, and Jack Citrin. 2007. "European Opinion about Immigration: The Role of Identities, Interests and Information." *British Journal of Political Science* 37(3):477–504.
- Simonsohn, Uri, Joseph P. Simmons, and Leif D. Nelson. 2015. "Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications."
- Stinchcombe, A. L. 1999. "A Treatise on Social Theory, Vol. 3: Applied Social Theory." *American Journal of Sociology* 104(48952):1241–1243.
- Textor, Johannes, Benito van der Zander, Mark S. Gilthorpe, Maciej Liśkiewicz, and George TH Ellison. 2016. "Robust Causal Inference Using Directed Acyclic Graphs: The R Package 'Dagitty.'" *International Journal of Epidemiology* 45(6):1887–94.
- Weber, Max. 2002. "Wissenschaft Als Beruf (1919)." Pp. 474–512 in *Max Weber Schriften 1894-1922*, edited by D. Kaesler. Stuttgart: Kröner.
- Wofle, Lee M. 2003. "The Introduction of Path Analysis to the Social Sciences, and Some Emergent Themes: An Annotated Bibliography." *Structural Equation Modeling: A Multidisciplinary Journal* 10(1):1–34.
- Wright, Sewall. 1934. "The Method of Path Coefficients." *The Annals of Mathematical Statistics* 5(3).

- Wu, Chunsen. 2019. *DAG: Stata Module to Provide Utilities for Directed Acyclic Graphs*. Statistical Software Components: Boston College Department of Economics.
- Young, Cristobal. 2015. "Sociologists Need to Be Better at Replication." *Orgtheory*. Retrieved ([link](#)).
- Young, Cristobal. 2018. "Model Uncertainty and the Crisis in Science." *Socius* 4:23

4 Relevance of sex, gender and/or diversity: Not applicable

5 Supplementary information on the research context

5.1 Ethical and/or legal aspects of the project

5.1.1 General ethical aspects

5.1.2 Descriptions of proposed investigations involving experiments on humans or human materials: Not applicable

5.1.3 Descriptions of proposed investigations experiments on animals: Not applicable

5.2 Data handling

I plan to centralize all data gained from the literature including our encoded causal models using a systems approach where everything is stored on a server and integrated programs can be used to access and interact with the data. This will all be publicly available, ideally through the infrastructure setup by the SPP, but can easily be done using the SOCIUM server in Bremen.

I plan to collect theoretical arguments from researchers in the field. These should all be made public; however, they do require consent forms for us to use and share their information.

What is of crucial importance to me is that the data and any goods developed by the project are available to all scientists and members of the public in the safest and easiest to understand and use manner as possible. This means having functional APIs, within-web browse data exploration features and porting to popular software such as Stata and R Studio.

5.3 Other information

6 People/collaborations/funding

6.1 Employment status information

Breznau, Nate. Postdoctoral Researcher and Principal Investigator. University of Bremen (66%), German Science Foundation (33%). Fixed-term until August 31st, 2022. Extension contingent on second round funding of the Sonderforschungsbereich 1342, "The Global Dynamics of Social Policy", or on this current application's success.

6.2 First-time proposal data: not applicable

6.3 Composition of the project group

Breznau, Nate, Dr. (see 6.1 above for details).

6.4 Researchers in Germany with whom you have agreed to cooperate on this project

Breiter, Andreas. Professor of Applied Computer Science at the University of Bremen. Given my basic knowledge of machine learning and developing algorithms for decision-making, I am pleased that Prof. Brieter will provide expert consultation and assistance in developing the complex machine tasks necessary to compare causal models, mine text data to identify particular features of interest in articles and assist with centralizing and economizing the desired systems for deployment across disciplines. I already work with Prof. Breiter in the development of WeSIS (Welfare State Information Systems) as part of the Collaborative Research Center, “The Global Dynamics of Social Policy” at the University of Bremen.

6.5 Researchers abroad with whom you have agreed to cooperate on this project

Freese, Jeremy. Professor of Sociology at Stanford University, United States. Prof. Freese has published extensively on replication and using meta-analysis to improve objectivity in social science. His work transcends disciplines and he has contributed to the meta-replication SCORE led by Brian Nosek. He is excited about my project idea and his expertise will be extremely useful as my project is developing a new method of meta-causal model comparison. He is also a Stata developer and his knowledge of software systems and development will be extremely useful. He is one of the more successful scholars in sociology and methods globally, thus his input and supervision on the project, in addition to publication collaboration, should be of great value.

6.6 Researchers with whom you have collaborated scientifically within the past three years

Maureen A. Eger. Umeå University, Sweden.
 Carola Hommerich. Sophia University, Japan.
 Herbert Obinger. University of Bremen, Germany.
 Carina Schmitt. University of Bremen, Germany.
 Eike M. Rinke. University of Leeds, United Kingdom.

6.7 Project-relevant cooperation with commercial enterprises

It is possible that in developing R Studio and Stata packages that the project collaborates with the developers of the base programs. This depends on the success of the project. For the computer systems planned for development, they will ideally be hosted in Munich as part of the infrastructure of proposed by the Principal Investigators of this SPP, or in Bremen as part of the Universities' infrastructure. The idea, is not to rely on private interests to develop our systems, to make them free and accessible to all scholars.

6.8 Project-relevant participation in commercial enterprises

6.9 Scientific equipment

The SOCIUM research center in Bremen has server capacity to facilitate large-scale applications. The hosts that run the virtualization software are two servers with two Intel Xeon E5-2687Wv4 CPUs with 256GB RAM each. They have storage with HDDs that amount to 54TB storage and 16TB SSD storage. The Stata/R server has two AMD EPYC 7351 CPUs, with a total of 32 cores (64 including hyper threading) and 512GB RAM. In short, the computer system application will have no problem running our project's needs for the first 36 months. However, I imagine that upon successful testing, the infrastructure for the computer system will be moved/developed within the