The Crowdsourced Replication Initiative: Investigating Immigration and Social Policy Preferences using Meta-Science. Executive Report.

Preprint · January 2019

DOI: 10.31235/osf.io/6j9qb

Nate Breznau
Universität Bremen
20 PUBLICATIONS **98** CITATIONS
SEE PROFILE

Alexander Wuttke
Universität Mannheim
13 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Alexander Greinert
Goethe-Universität Frankfurt am Main
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Amélie Godefroidt
KU Leuven
9 PUBLICATIONS **11** CITATIONS
SEE PROFILE

The Crowdsourced Replication Initiative: Investigating Immigration and Social Policy Preferences

‹ Executive Report ›

Principal Investigators:

Nate Breznau, University of Bremen breznau.nate@gmail.com

Eike Mark Rinke, University of Leeds E.M.Rinke@leeds.ac.uk

Alexander Wuttke, University of Mannheim alexander.wuttke@uni-mannheim.de

Version 1.0.2, 1/29/2019

Research Participant Co-Authors:

Muna Adem, Jule Adriaans, Amalia Alvarez-Benjumea, Henrik Andersen, Daniel Auer, Flavio Azevedo, Oke Bahnsen, Dave Balzer, Paul Bauer, Gerrit Bauer, Markus Baumann, Sharon Baute, Verena Benoit, Julian Bernauer, Carl Berning, Anna Berthold, Felix S. Bethke, Thomas Biegert, Katharina Blinzler, Johannes N. Blumenberg, Licia Bobzien, Andrea Bohman, Thijs Bol, Amie Bostic, Zuzanna Brzozowska, Katharina Burgdorf, Kaspar Burger, Kathrin Busch, Juan Castillo, Nathan Chan, Pablo Christmann, Roxanne Connelly, Christian Czymara, Elena Damian, Alejandro Ecker, Achim Edelmann, Maureen A. Eger, Simon Ellerbrock, Anna Forke, Andrea Forster, Chris Gaasendam, Konstantin Gavras, Vernon Gayle, Theresa Gessler, Timo Gnambs, Amélie Godefroidt, Alexander Greinert, Max Grömping, Martin Groß, Stefan Gruber, Tobias Gummer, Andreas Hadjar, Jan Paul Heisig, Sebastian Hellmeier, Stefanie Heyne, Magdalena Hirsch, Mikael Hjerm, Oshrat Hochman, Jan H. Höffler, Andreas Hövermann, Sophia Hunger, Christian Hunkler, Nora Huth, Zsofia Ignacz, Laura Jacobs, Jannes Jacobsen, Bastian Jaeger, Sebastian Jungkunz, Nils Jungmann, Mathias Kauff, Manuel Kleinert, Julia Klinger, Jan-Philipp Kolb, Marta Kołczyńska, John Kuk, Katharina Kunißen, Dafina Kurti, Philipp Lersch, Lea-Maria Löbel, Philipp Lutscher, Matthias Mader, Joan Madia, Natalia Malancu, Luis Maldonado, Helge Marahrens, Nicole Martin, Paul Martinez, Jochen Mayerl, Oscar J. Mayorga, Patricia McManus, Cecil Meeusen, Daniel Meierrieks, Jonathan Mellon, Friedolin Merhout, Samuel Merk, Daniel Meyer, Jonathan Mijs, Cristobal Moya, Marcel Neunhoeffer, Daniel Nüst, Olav Nygård, Fabian Ochsenfeld, Gunnar Otte, Anna Pechenkina, Christopher Prosser, Louis Raes, Kevin Ralston, Miguel Ramos, Frank Reichert, Leticia Rettore Micheli, Arne Roets, Jonathan Rogers, Guido Ropers, Robin Samuel, Gergor Sand, Constanza Sanhueza Petrarca, Ariela Schachter, Merlin Schaeffer, David Schieferdecker, Elmar Schlueter, Katja Schmidt, Regine Schmidt, Alexander Schmidt-Catran, Claudia Schmiedeberg, Jürgen Schneider, Martijn Schoonvelde, Julia Schulte-Cloos, Sandy Schumann, Reinhard Schunck, Jürgen Schupp, Julian Seuring, Henning Silber, Willem Sleegers, Nico Sonntag, Alexander Staudt, Nadia Steiber, Nils Steiner, Sebastian Sternberg, Dieter Stiers, Dragana Stojmenovska, Nora Storz, Erich Striessnig, Anne-Kathrin Stroppe, Janna Teltemann, Andrey Tibajev, Brian Tung, Giacomo Vagni, Jasper Van Assche, Metavan der Linden, Jolanda van der Noll, Arno Van Hootegem, Stefan Vogtenhuber, Bogdan Voicu, Fieke Wagemans, Kyle Wagner, Nadja Wehl, Hannah Werner, Brenton Wiernik, Fabian Winter, Christof Wolf, Nan Zhang, Conrad Ziller, Björn Zakula, Stefan Zins and Tomasz Żółtak[1].

[1] We also recognize and are grateful for the partial contributions of Esra Akdeniz, Ling Bai, Christina Eder, Danilo Freire, Verena Halbherr, Sabine Israel, Jennifer Kanjana, Salman Khan, Sayak Khatua, Robin Lee, David Liu, Robert McDonnell, Jordan Suchow, Cary Wu and Yuki Yamada. Six of the co-authors' team had not yet submitted

PRINCIPAL INVESTIGATORS

Nate Breznau is a postdoctoral researcher at the University of Bremen. He obtained his PhD in Sociology at the Bremen International Graduate School of Social Sciences. His research focuses on public opinion, social policy, immigration, survey data and social inequality. He works on the global history of social policy at the Comparative Research Center "The Global Dynamics of Social Policy" (SFB 1342) and is PI of the German Science Foundation project, "The Reciprocal Relationship of Public Opinion and Social Policy". His entry into open science came with his attempt to replicate intransparent research during his dissertation.

Eike Mark Rinke is a lecturer in Politics and Media at the University of Leeds. He obtained his Dr. phil. in Media and Communication Studies at the University of Mannheim. His research focuses on the empirical study of normative aspects of political communication with a particular interest in mediated and face-to-face forms of political deliberation and discussion, journalism and news cultures in a comparative perspective. His entry into open science came with his observing the often wide gap between the ideal and actual worlds of scientific practice.

Alexander Wuttke is a political psychologist at the University of Mannheim. He examines individuals' perceptions and evaluations of the political environment in which they are embedded and how these orientations shape motivation for political engagement. In his dissertation, he synthesizes psychological theories to investigate the origins of political motivation.

the expansion as of publication of this pre-print therefore they were not included as authors. If they submit by the time of our journal publication they will be included in the article.

FOREWORD

We three PIs developed the ideas behind this study while working as researchers at the Mannheim Centre for European Social Research (MZES) at the University of Mannheim, Germany. In addition to being a topic of substantive interest, part of the groundwork for this study came from Breznau's previous experience in replication and identifying researcher variability (Breznau 2015, 2016). The main inspiration for the methods behind the study came from Silberzahn et al.'s pioneering large-scale crowdsourced study of football data (2015). Overall motivation for the project came from the open science movement and our intention to ferry it from psychology into social sciences such as political science, communications and sociology. It also came from our organization of the *MZES Open Social Science Conference* (Jan. 25-27th, 2019) and from our own intentions to practice better, more ethical, transparent, and useful science[2].

[2] We are grateful to David Brady and Ryan Finnigan for their cooperation and Timo Dobbrick for his support.

TABLE OF CONTENTS

1 OVERVIEW

In an era of mass migration, a wave of social science, populist parties and social movements raise concerns over the future of immigration-destination societies. A burning question is what impact immigration and immigration-related events have on policy and social solidarity. Comparative cross-national research on immigration and policy preferences of the public mostly uses secondary data, and the findings go in different directions. There is great potential researcher bias in both selective model reporting and lack of replicability in this area. Moreover, at the macro-level, the heterogeneity of countries obscures attempts to clearly define data-generating models. It is easy to run every possible model in the secondary data and report only those that appear interesting or significant.

This project employs crowdsourcing methods to address these issues. These methods follow the open science movement as a response to some of the crises in science. Crowdsourcing in the social sciences is a rather new and unexplored method. It draws on replication, deliberation, meta-analysis and harnessing the power of many minds at once. Therefore, the Crowdsourced Replication Initiative (CRI) carries two main goals, (a) to better investigate the linkage between immigration and social policy preferences across countries, and (b) to develop crowdsourcing as a social science method. This executive report provides short reviews of the area of social policy preferences and immigration, and the methods and impetus behind crowdsourcing. It details the methodological process we followed conducting the CRI and provides some descriptive statistics. The project has three planned papers as follows. Paper I will include all participant co-authors; II and III just the PIs. Readers may follow the progress of these papers and find all details about the CRI on its Open Science Framework (OSF) project page[3].

I. "**Does Immigration Undermine Social Policy? A Crowdsourced Re-Investigation of Public Preferences across Mass Migration Destinations**". The main findings of the project presenting a replication and extension based on original research of Brady and Finnigan (2014) titled, "Does Immigration Undermine Social Policy Preferences?".

II. "**Deliberative Research: Can Reasoned Debate Improve the Scientific Process?**". Based on an experimental condition, findings on how deliberation potentially improves the method of crowdsourcing, in particular in its impact on the research process.

III. "**How Reliable Are Replications? Measuring Routine Researcher Variability in Macro-Comparative Secondary Data Analyses**". Based on an experimental condition testing two ideas: (1) to identify and quantify error associated with the variability in results from researchers working with the same data and models, and (2) to test if this variability is larger when the original study is less transparent with its methods and results.

[3] https://osf.io/bs46f/

1.1 Immigration and Social Policy Preferences

Increasing rates of immigrants and their offspring generate reactive opinions, social movements and policies across rich democratic welfare states. These are the states that constitute the primary destination societies for immigrants who are both selectively mobile or forcibly displaced. The effects spill over into other societies that have little to no immigration visible for example in movements among Eastern European societies in response to events in Western Europe and changes at the level of European governance. Somewhat similarly, just the fact that the percentage of foreign-born persons increased from 1 to 2 percent in a society such as Japan or Korea causes growing concerns among natives, despite very few immigrants in the population overall.

A fundamental condition shared by destination societies is the need for immigrants. Either immigration or unprecedented economic growth *must* take place to support the aging populations amongst the rich democratic welfare states. In the 'golden age' of welfare state security, the old-age dependency ratio was 4-to-1. Four workers for every pensioner. By 2050, sources predict this rate somewhere under 2-to-1 in most aging societies and as low as 1-to-1 among hyper aged societies such as Japan (Breznau 2018). There simply are not enough young, healthy workers across the occupational spectrum. In particular, not enough to support the consumption and welfare needs of aging populations. Economic growth at rates similar to the post-war period are unlikely, they decreased steadily on average for fifty years. If there is continuation of reduced or flat growth then immigration is necessary to prevent collapse of social welfare, specifically for those who need it most. Interestingly, some members of the public and scholars predict that social welfare might collapse because of too much immigration, not a lack of it.

There are various theories suggesting that increased presence of foreign-born persons in destination societies lead members of those societies to become less supportive of social welfare promoting policies (see reviews in Brady and Finnigan 2014; and Schmidt-Catran and Spies 2016). According to variants of social identity theory, national group boundaries come into play for native members of societies under increasing immigration rates (Hjerm 1998; Semyonov, Raijman, and Gorodzeisky 2006). Native group boundaries become more salient. This relates to the rather generic concept of conflict theory where groups compete for scarce resources and that immigrants constitute an out-group with which natives do not want to share their resources (Schneider 2008). Interestingly, game theoretic experiments assigning group boundaries randomly to participants led to the discovery that policies (e.g., resource sharing) that seem to benefit the out-group, even if it also benefiting the in-group, are not preferable (see reviews in Eger 2010; and Mitchell 2018). In other words, in-group members are often willing to take away resources from their own group to prevent a perceived out-group from gaining access to them.

Many other theoretical arguments suggest that the impact of immigration depends subnational dynamics. Group threat theory suggests immigration may lead to increased support for social policy to protect workers and their families against potential job loss resulting from immigration. But this effect is only likely in industries or sectors with more immigrant workers (Alt and Iversen 2017; Burgoon 2014). There is a distinction between conflict and contact. For example, higher educated persons are more likely to see the need for immigrants, to value cultural diversity and express egalitarian principles (Breznau and Eger 2016). In part because they may have secure jobs less affected by immigrant competition or they have higher levels of social capital, in particular weak-ties, and can easily find jobs in case of competitive job losses (McLaren 2017). This suggests immigration has positive impacts on these selective persons.

There are also researchers arguing that institutional policy settings condition the impact of immigration, with more egalitarian, universal and immigrant affirmative policies less likely to react negatively (Kesler and Bloemraad 2010). In the massive literature on the subject there are arguments suggesting that immigration could increase support for social policy, or have no impact. Not only do findings support several conflicting conclusions about the link between immigration and social policy preferences, but other research challenges the entire enterprise given that research can depend entirely on seemingly trivial coding or indexing choices at the country or policy-levels measures stand on shaky empirical grounds as the coding and indexing of country-level variables itself can produce great variation in empirical results (Bjerre, Römer, and Zobel 2019; Green-Pedersen 2004; Kunißen 2018).

Arguably, theories of immigration and the sustainability of social solidarity, social welfare or the state itself are too ‘big’. There are various sub-national and intra-individual processes that make each policy, policy discussion and public opinion at any given time and place unique. A grand theory may not cover all the complexities. In other words, there is too much confounding unobserved heterogeneity. However, we remain focused on the ‘big’ picture for one primary reason. *What if immigration, generally speaking, does undermine social policy preferences, and thus enables social policy retrenchment to the point that inequality destroys societies because of revolution or intra-state violence?* The negative impact of inequality 21st Century societies seems non-negligible (Pickett and Wilkinson 2015; Therborn 2013). Although Brady and Finnigan (2014) concluded no impact of immigration on social policy preferences in any systematic way, we argue that a potential threat to societies this large warrants continued scrutiny. Moreover, recent advances in knowledge about scientific objectivity and reliability suggest extra scrutiny of all research[4]. As we will see in the next section we should particularly scrutinize macro-comparative secondary data analysis.

[4] As we discuss in our methods section, there is nothing particularly alarming about Brady and Finnigan’s study. It is very much a standard form of research in this area. We did not single it out for any reason other than its visibility/impact, transparency and case coverage.

1.2 The Crisis of Social Science & Macro-Comparative Secondary Data Analysis

The area of immigration and social policy research, like most any interdisciplinary field dependent on secondary data from macro-comparative surveys, is an ideal setting to discuss some persistent problems of science. By now, members of the public, policymakers and a growing number of social scientists are aware of the crisis of science (Delfanti 2010; Saltelli and Funtowicz 2017; Wuttke 2018). Experimental sciences were hit particularly hard by this crisis, but statistics in general and secondary data analysis in the social sciences suffer as well. For example p-hacking, HARKing and the researcher degrees of freedom problem in the 'garden of forking paths' face all scientists (Gelman and Loken 2014; Munafò et al. 2017). There is also a replication or reproducibility crisis that brought forward much media attention in addition to the high profile data faking scandals (Baker 2016; Ioannidis 2005).

Replication, according to Clemens (2015:327), "estimates parameters drawn from the same sampling distribution as those in the original study." In macro-comparative research, usually only replications involving the same survey data are possible, thus estimates can only be drawn from the exact same sample, instead of the same sampling distribution. This limits replications to verifiability and robustness testing (Freese and Peterson 2017), because they lack the means to draw a new international survey sample from the same population.

As the open science movement made us aware: In macro-comparative research with secondary data, researchers can and will run up to billions of models (Orben and Przybylski 2019; Young 2018). Thus, by default secondary macro-comparative research in this area posits a spectrum of data-generating hypotheses that are nearly infinite. Each different model represents one or several potential theory(ies) of the data-generating process. Whether researchers acknowledge this or not, it is implicit in specifying Y and X variables and their relationships. However, even given the most appropriate model design choices for expressing the data-generating model, results might still differ. Even using the same data and same models we may observe researcher variability. Although some argue for the utility of millions of models (Muñoz and Young 2018), in crowdsourced research the set of models are not a simulation but a set of *real models derived from practicing social scientists*. No matter how strong the statistical methods, data itself cannot produce the underlying causal model, this requires rational construction of theory, exclusion restrictions, attention to confounding and counterfactuals (Pearl 2010)[5]. Thus, in a crowdsourced project, we have several plausible data-generating models rather than all possible models, plausible or not. Moreover, as the crowdsourced replication starts with raw data we additionally gain a real-life (rather

[5] Note that those supportive of machine learning might disagree here, but we do not support this position. Machine learning can make the most efficient possible predictions. It requires social thought to explain what caused the outcomes behind these predictions.

than simulated) multiverse of data variations that result from seemingly benign coding and cleaning choices (Breznau 2016; Steegen et al. 2016).

Given the sensitivity of small-N macro-comparative studies and that researchers might report only those models out of the thousands that gave the results they sought, reliance on a single replication study (or the original), even if it is deemed to perfectly reflect the underlying causal theory, is a flimsy means for concluding whether a study is verifiable. With a pool of researchers independently replicating a study we develop far more confidence in the results. As our preliminary power analyses conclude, it takes several independent replications to ensure that a majority come to the correct conclusions (Breznau, Rinke, and Wuttke 2018).

In this project we wanted to test if crowdsourcing was possible under these secondary macro-comparative conditions. In the Silberzahn et al. (2018) study, the data were never before seen football red card data. The participants were not experts in football analysis. The entire project filled exogeneity criteria such that the researchers likely had little preconceived bias about the study – both when selected, as they were unaware of the topic, and during the research process, as these were a type of data they never saw before. In our case, both conditions are opposite. Given the nature of the study we found it necessary to reveal some of the details during the call for researchers. Moreover, there is a preexisting study and a huge body of literature involved. Finally, many researchers previously worked with the data we asked them to analyze. Thus, a goal of ours was to test if and how crowdsourcing could work in the area of replication and macro-comparative secondary data analysis.

1.3 Crowdsourced Learning – Methodological Innovations & Experimental Designs

Crowdsourcing as a term is rather young. It developed in reference to the capacity of any individual internet user to contribute to the content and outcomes of online platforms and discussions such as Wikipedia (Zhao and Zhu 2014). As a scientific practice, however, its roots go back to open engineering competitions in the 1700s (Sobel 2007). Through such competitions, researchers, engineers and business developers overcame problems that were previously insurmountable. This was done by simply asking the public or many specialists to use their computing power or problem-solving skills; problems as varied as improving climate change predictions, cracking ciphers and cancer research can be improved in this way (Howe 2006).

Given our conditions of topic and study selection, only methodological example on which to develop our crowdsourcing methods comes from the Silberzahn et al (2018) study[6]. Using their work as a basic shell for our project we developed methods to improve crowdsourcing and test various hypotheses

[6] And their detailed catalog of methods on the project's OSF page https://osf.io/gvm2z/

alongside our main hypothesis about immigration and social policy preferences. Therefore, we planned two experiments. One during the replication phase to assess variability in replications, and the other in between replication and expansion involving an experimental deliberation.

1.3.1 Researcher variability experiment

We hypothesize that there is researcher variability in replications even when the original study is fully transparent (Breznau et al. 2018). Therefore, we wanted to test the limits of this variability and identify how replicators might have more or less replication success depending on the transparency of an original author's study. To do this we randomly divided our researcher sample into an original group and an opaque group. The original received all materials including code from the original study while the opaque group received an anonymized version of the study, with only descriptive rather than numeric results and no code.

1.3.2 Deliberation experiment

We hypothesize that active deliberation among researchers will lead to increased affect and learning during the crowdsourced project (Wuttke, Rinke, and Nate Breznau 2018). We also considered the possibility that researchers might change their research designs if they participate in deliberation. Therefore, after the replication we did a second randomization of all participants with half continuing to work on their research designs independently and the other half deliberating their research design choices in an online platform.

1.3.3 Researcher panel survey

We speculated that researchers' own qualities might influence their replications, research designs and expansion results. Therefore, we surveyed researchers on both objective criteria, such as experience with methods and the substantive topic, and subjective criteria, such as their own beliefs about the hypothesis and immigration in general. In addition, we asked them questions about their time commitment, constraints they faced and some other feedback about the process of crowdsourcing. We conducted 4 waves throughout the CRI allowing us to field a core questionnaire and determine if participation in the CRI or either experimental condition might alter subjective perceptions and experiences.

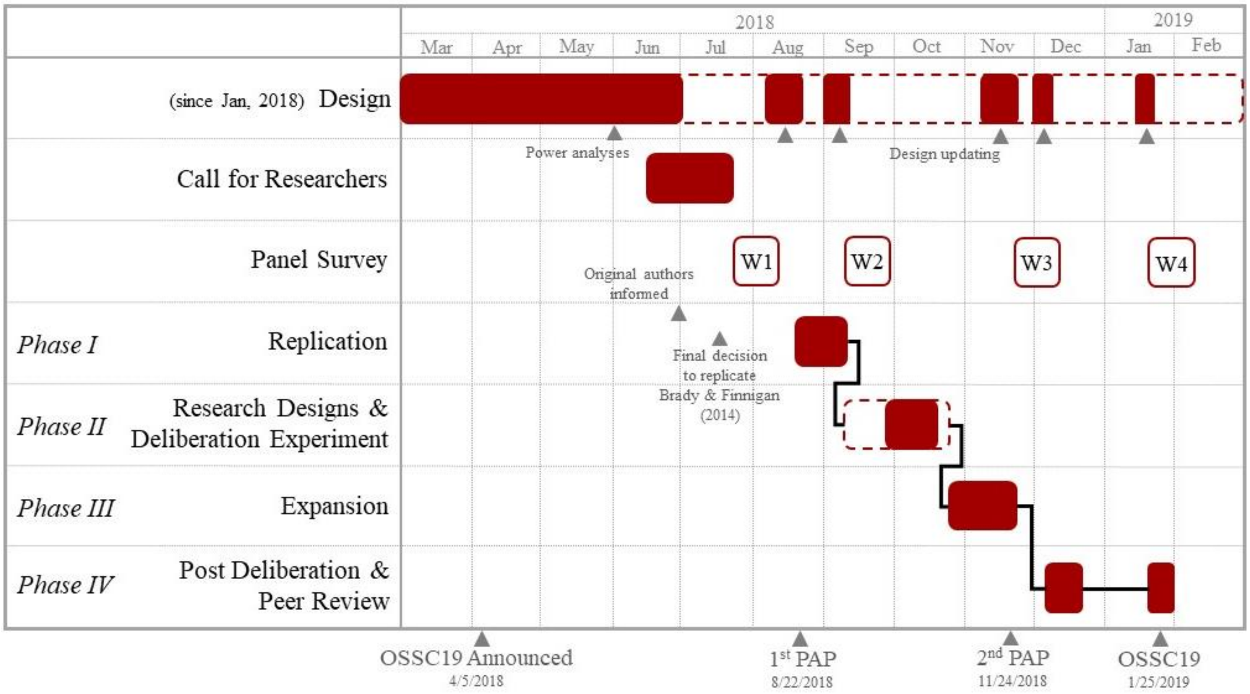
2 METHODS & SELECTED DESCRIPTIVE STATISTICS

We employ crowdsourced replication to test the hypothesis that immigration undermines support for social policy across rich democracies, also known as a ‘many analysts’ approach (Silberzahn et al. 2018). This project involves a replication and then an expansion of Brady and Finnigan’s (2014) study (“B&F” throughout this section) titled “Does Immigration Undermine Public Support for Social Policy?”. Our study included 106 research teams initially and after dropout resulted in 88. Teams of researchers or solo researchers applied to take part with the condition that they have some multilevel regression skills. The study called for researchers from all disciplines, at all levels of research, including professors, postdocs, PhD students, lecturers and analysts working in non-academic sectors, although the mention of social science and the topic of immigration, social policy and public opinion likely selected mostly a pool of researchers with at least some expertise in the area. We conducted initial power analyses that revealed we would need a minimum of 40 participants to do a replication treatment and a minimum of 60 to do a deliberation treatment. We elected to hold the call for researchers before completing our pre-registration plan as we had no idea how many we would get, and our methods depended on this (Breznau et al. 2018).

Figure 1 provides a timeline of the CRI from its inception in January, 2018 during discussions about how to do open science after Wuttke invited Breznau and Rinke to collaborate in the organization of the *MZES Open Social Science Conference 2019* (OSSC19). The details of communications with participants, power analyses in our registered pre-analysis plans (PAPs) and survey questionnaires are all available on the CRI’s Open Science Framework (OSF) project repository[7].

[7] https://osf.io/bs46f/

Figure 1. The Crowdsourced Replication Initiative Timeline



2.1 *Replication Selection*

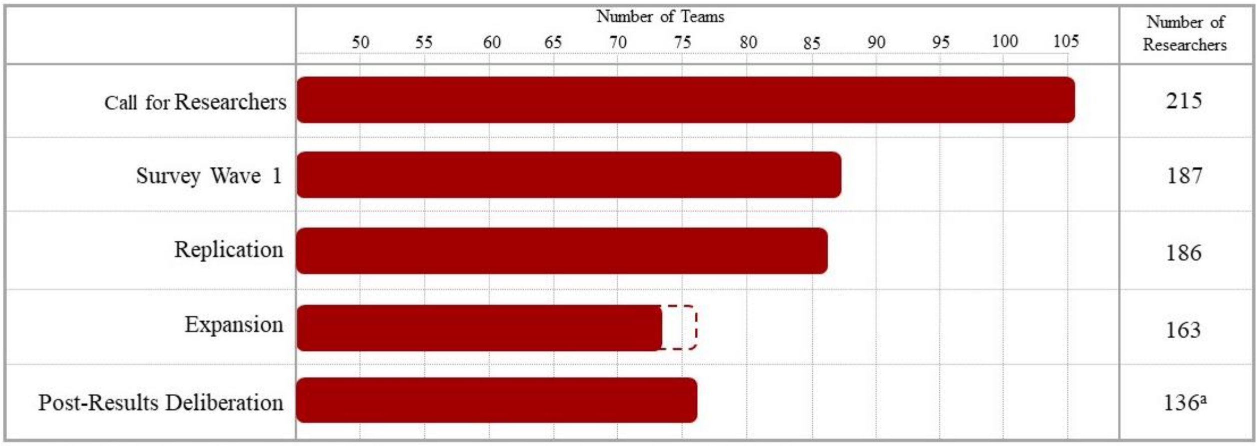
The building blocks for crowdsourcing are already present in B&F's analytical framework. They used pooled regressions and two-way fixed-effects models, both analyzing stock of foreign-born and net migration separately by model type. They run extensive alternative model configurations adding in a different dependent variable (asking about spending) and several different independent variables (Multiculturalism Policy Index, regime, non-Western migration). They use all of their models together to conclude that their results do not support the hypothesis that immigration undermines social policy preferences. To make replications feasible we focus only on their two-way fixed effects models. These are the only models that offer any insight whatsoever into change over time at the country-level, the level where we would expect to observe an effect if one truly existed (Tables 4 and 5 in B&F's original work). In other words, these are the only models that attempt to test if changes in stock or net migration of foreign persons shapes attitudes within countries over time. All of these alternative model specifications suggests that they or the reviewers of their work were not settled on one particular 'true model' of the data-generating process. As such, their work simply reflects the state of the art in this research area which has all kinds of model specifications as we expect in crowdsourcing.

Another advantage of the B&F study is transparency. Unlike many studies published in top social science journals like the *American Sociological Review*, they provided supplementary materials online to demonstrate the results of their alternative model specifications and Stata code. A decisive factor for us selecting this study was its prominence in one of sociology's, if not social sciences in general, highest ranked journals. In the process of selection, we contacted B&F and they were supportive of our potential plans to replicate their work. This made for a communicative and open process leading us to feel very comfortable in our research efforts. In a pretest, two of the PIs of the project could independently replicate their work suggesting that the study is verifiable.

2.2 Researcher Participants

We had 216 researchers in 106 teams from 26 countries in 5 continents respond to our call for researchers as of its closing on July 27th, 2018. Roughly 44% were female[8]. Figure 2. Gives the rates of participation throughout the course of the CRI.

Figure 2. CRI Participation Rate



a 'Full Participation' required completion of the Replication and Expansion, not necessarily the Deliberation

The call for researchers promised co-authorship on the final published paper, similar to the Silberzahn et al (2018) study. For us this was an essential component to reduce if not remove publication or findings biases. We wanted to be clear that simply doing the tasks we assigned qualified them as equal co-

[8] Team registration required only one person to respond at that time, thus we have data only for the registrant, not all team members.

researchers and co-authors, they needed not produce anything special, significant or 'groundbreaking', only solid work.

2.3 Phase One – Replication

In macro-comparative research, usually only replications involving the same survey data are possible, thus estimates can only be drawn from the exact same sample, instead of the same sampling distribution. This limits replications to *verifiability* and *robustness* testing, because they lack the means to draw a new international survey sample from the same population. In a perfectly transparent world, verifiability should be a matter of (a) checking researchers' software programming code to determine if it indeed produces the reported results, and (b) determining if the methods reflect what the researchers claim they did in their published study. Currently, the world of macro-comparative research with secondary data is far from transparent[9] leading to many limitations in attempts at verifiability. Moreover, the experience of journals requiring code with their submissions demonstrates that code alone is not necessarily enough to replicate (Eubank 2016)[10]. Yet, our hypothesis is that even with full transparency, verifiability may not be perfectly reliable due to researcher variability.

Prior to carrying out the research reported herein, we preformed power analyses to determine how many replications might be necessary to identify a true effect given routine researcher variability. If a standard rate of failed replications due to researcher variability is 30%, at least 15 replicators would be necessary to recover the correct replication of the original given no other sources of confounding (see Table 2 in Breznau et al. 2018). To perform our testing, we set up an experiment varying conditions which might point toward the size of researcher variability, generalizability of our findings, and test for a difference between replications conditioning on the transparency of the original study. We have two replication conditions.

In the first group, labeled *original version*, we assign teams to assess the verifiability of the prominent B&F macro-comparative study. This group has minimal research design decisions to make, theoretically none. They engage in checking the work and results of the B&F study that uses *International Social Survey Program* (ISSP) data based on questions about the government's responsibility to provide various policies targeting social welfare of the population. The study aggregated and regressed these on immigration indicators of percent foreign-born and net migration at the country-level, plus some independent variables at both the country and individual-levels. As the original study is very transparent

[9] Based on results of research by Elena Damian, Bart Meuleman and Wim van Oorschot presented at the OSSC19, "Evaluation and Replicability Transparency in Cross-National Survey Research: Quality of Reporting".

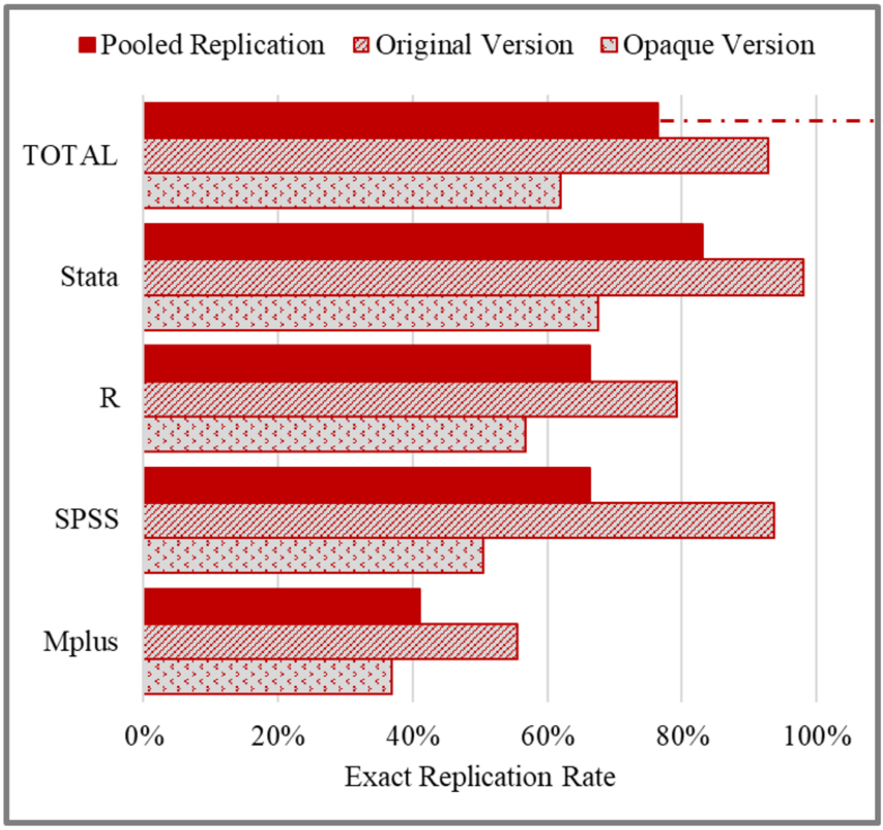
[10] Nicole Janz, May 4th, 2015, "Leading Journal Verifies Articles before Publication – so far, All Replications Failed" https://politicalsciencereplication.wordpress.com/2015/05/04/leading-journal-verifies-articles-before-publication-so-far-all-replications-failed/

with the authors sharing their analytical code (for the statistical software *Stata*) and country-level data, it is a least-likely case to find variation in outcomes. This allows us to conduct a conservative test. Any variations within this experimental group should constitute routine researcher variability except for a set of control variables we measure.

We gave the second group a slightly artificial treatment. They replicate a derivation of the B&F study altered by us to render it anonymous and less transparent, labeled *opaque version*. This provides a controlled experiment to simulate what might magnify researcher variability in robustness testing replications where there are more choices and thus more chances for routine variability to enter the research. Although the researchers are given only the task of assessing verifiability, we simulate additional 'robustness-like' choices by offering them limited information about the original study, forcing them to make 'tougher' choices in how to replicate it. To do this we re-worded the study, gave no numeric results and provided no code to the replicators in this experimental group. This opaque version also restricts a different kind of choice because the replicators do not have numeric results to look at. We theorize that having results might lead the replicators to continue adjusting their models until they arrive at the results of the original study, a phenomenon akin to confirmation bias, thus giving them a motivation to make changes (i.e., choices) and a specific goal in making these changes (Munafò et al. 2017).

Replicators reported odds-ratios following the estimates reported by B&F. Each team reported between 40 and 48 odds-ratios depending on experimental condition, resulting from several models and inclusive of both the immigration test variables and the country-level plus the independent variables of social spending and employment rate. We considered an odds-ratio to be an "exact" replication if it was within <0.01 of the original effect; this allows for rounding error. Figure 3 reports the ratio of exact replications by usage of software types and experimental group.

Figure 3. Phase One Replication Software Usage & Results by Group



Exact	Total
2,626	**3,438**
1,493	1,609
1,133	1,829
1,872	2,255
1,119	1,141
753	1,114
588	886
299	377
289	509
116	175
60	64
56	111
50	122
15	27
35	95

Identification of routine researcher variability is conditional on identification of non-routine researcher variability. Non-routine variability is due to specific choices such as changing model features, adding extra steps or major coding mistakes (i.e., sloppy science), see Table 1. We must identify this and adjust the replicators' results to identify the magnitude of routine researcher variability – that which appears out of idiosyncratic features or undeliberate actions of the researchers. We employ two methods to achieve this. The first is that in the course of the larger crowdsourcing project we surveyed participants on some background variables, familiarity with the original study, subjective positions regarding the topic at hand and their overall skill with methods and multilevel methods in particular. We also identify the discipline of study and the type of software used. We test these variables on the replicator results to see if any show consistent biasing patterns and see what variability remains after conditioning on these variables in a regression to the mean of replicators results.

Next, we will code each replicators' results into three dependent variables. Exact replication refers to precise recovery of the numerical estimates provided by the original study. We have a continuous variable representing the absolute difference between the replicators' results and the original study’s odds-ratios multiplied by negative 1 resulting in a continuous variable where numbers with greater negative values represent results further form the original (i.e., positive coding where higher values approach an exact replication). Finally, we have the researcher’s own subjective conclusion about the replication where we asked them to conclude whether they replicated the results of the original study, yes=1, no=0.

We establish grounded criteria for replicators’ decisions to rate a study as verifiable or not and in arriving at their numerical results. Table 1 lists what type of variability results we understand as caused by the object of each independent variable, as developed in our pre-registered report (Breznau et al. 2018:Table 3). In doing so we will remove what we determine to be non-routine researcher variability. The variance that remains after controlling for this should be evidence of routine researcher variability. To make our estimates as conservative as possible we also exclude some or all of variability that might be a product of both routine and non-routine researcher variability. Using the size of our estimates of routine researcher variability plan simulations to demonstrate how it could impact social science research.

Table 1. Distinguishing Two Forms of Researcher Variability

	Routine	Non-Routine
Mistakes	Minor coding or reporting mistakes - idiosyncratic events or traits	Major coding or reporting mistakes - 'sloppy science'
Expertise of the researcher	Should reduce mistakes	Should reduce mistakes
Modeling	Unintentional features of a model generated in the construction phase	All decisions in constructing a formal or statistical model
Software	A researcher's standard software type and version	Exception: use or learning of a non-routine alternative software
	The defaults of the typical software	--
Extra steps	Unintentionally adding or altering an analysis, something not metioned in an original study being replicated for example	Exception: when a researcher adds steps to a model intentionally to produce results (like p-hacking)
Access	Institutional or personal limitations in access to software, data or other necessary resources	--
'Random' Error	Variability that cannot be controlled	Variability that can be accounted for, debated or explained by rational choices and intentional actions
	Specific to Replication Research	
Quality / Transparency of materials	Forces researchers to make more choices, introducing more opportunities for routine error	Forces researchers to make more choices, introducing more opportunities for non-routine error
Variability error is endogenous with original study	No. Exists in any replication attempt independent of original study	Yes. Is more likely when original study is controversial or lacks transparency

2.4 Phase Two – Research Designs & Deliberation

Most academic research takes place in a small-scale research style. Two widespread features of this style stand out: One is a centralized decision-maker. Many crowdsourced studies rely on choices and directions emanating from a central researcher or institution. The other is that - even if part of a real or imagined "crowd collective" working on a specific research problem - researchers and research teams usually work in isolation to develop their (partial) solutions. Decentral, well-integrated deliberation, on the other hand, has remained a largely untapped resource for crowdsourced academic research, much less academic research in general. When researchers give feedback to and deliberate about each other's research practices the entire process might make quantum leaps forward in terms of efficiency and effectiveness, pushing it closer to the long-ago espoused ideals of science as an open and widely collaborative process (Merton 1942).

However, once science leaves the familiar confines of vertically organized research and embraces research in larger teams integrated through deliberative interaction, new problems and unresolved questions arise. Therefore, as a means of improving both crowdsourcing and social science in general we conducted an experiment to gauge the impact of a deliberative crowdsourced process on the social and epistemic outcomes of our large-scale research project. Specifically, this experiment investigates how organized, reasoned debate among researchers affects the quality of research outcomes compared to the established model of atomistic organization without communication. Besides the practical relevance of this study for the design of future academic projects, this experimental design also speaks to meta-scientific and socio-scientific questions: It shows how organizational research contexts shape the outcomes of social inquiry.

This experiment depended on acquiring a large enough sample for its execution. Looking at the pool of researchers that signed up for the CRI we determined that we can detect treatment (deliberation) effects with an appropriate level of statistical power if the observed effect size is (Cohen's *d)* > 0.6 when conducting team-level analyses and > 0.4 when conducting individual-level analyses (Wuttke et al. 2018). We compare two experimental groups: participants who did not engage in any deliberation with those who participated in the deliberation. We aim to hold all other elements of the research process constant between both groups. Therefore, the online deliberations of the experimental group should be the only cause of statistically significant differences between the groups at the set level of alpha, as opposed to unobserved heterogeneity.

Our working hypothesis is that deliberation causes changes in researcher understandings, preferences, and subjective experiences relative to the research process. As a downstream consequence, we expect the individual-level effect to contribute to differences in the final outcomes of research conducted by those randomly assigned to deliberate versus those randomly assigned to be in a non-deliberating control group. We designate this as a "working" hypothesis because the expected insights into the effects of deliberation in a crowdsourced research context will, in part, be grounded in the subjective experiences of the deliberation participants and arrived at only through qualitative reconstructions. To our knowledge, this is the first experimental social science study to test how deliberation may or may not improve a large-scale crowdsourced research project.

Our hypotheses come in two formats:

Perception of Process Hypothesis: The deliberation experience had a positive impact on how participants perceived the research process.

Research Outcome Hypothesis: The deliberation experience had a positive impact on the quality or form of the research outcomes in four postulated ways:

1. *Reversibility*: Self-reported changes of mind based on responses to survey questions asking for degree of subjective change experience (continuous scale), corroborated with experimental evidence from hand-coded comparisons of the original and the revised research designs developed by researchers in the treatment (deliberation) group during the crowdsourced research process.
2. *Argument repertoire:* number of reasons participants can provide when prompted to justify their own and others' final research design proposal, based on human-coding of responses to open question (see Cappella, Price, & Nir, 2002).
3. *Quality of proposed final research design* as rated by the participants. Mean scores for each research design evaluation, compared between deliberation group and control group.
4. *Perceived success* of the crowdsourced research project, measured after conducting the research in Survey Wave 3 (see OSF Project website). Mean scores on these measures are expected to be higher (more positive) for the deliberation group.

We collect descriptive evidence on the deliberation process as it unfolded using an online deliberation platform Kialo. In designing the optimal deliberation experience, we learn from previous deliberation experiments (Escobar, 2011). Because the experiment's participants are spread around the globe, we make use of online communication. Therefore, we make particular use of the insights learned from previous deliberation experiments conducted online (Davies and Chandler, 2012; Friess and Eilders, 2015).

Given the importance of usability (Towne and Herbsleb, 2012), we selected an online communication platform that was developed using scientific insight with deliberative principles in mind. Kialo has three principle benefits. The first is that is facilitates short, written communications. Research suggests that messages that are too long will frustrate or perhaps be ignored by participants. The second is that it organizes communications into a knowledge tree. For each position or argument about a given topic, participants can make comments and these are branches that attach only to the original position. Moreover, comment-specific sub-comments attach only to the higher level comment. This provides an easy means for participants to view only the relevant communications when deliberating over a specific point. Finally, Kialo allows for voting. As deliberative processes over contentious or complex phenomena rarely reach consensus, in many deliberations this is explicitly not a goal, the process of voting provides a means of assessing ‘participant opinion’ or the most important points. This is especially useful as it quantifies the process of deliberation allowing for comparison of outcomes across experimental groups with minimal subjectivity.

In designing the deliberation process on Kialo, we take best-practice recommendations into account as suggested in reviews of previous online deliberation experiments (Davies and Chandler, 2012; Friess and Eilders, 2015). We invited all participants to Kialo and provided them with an instructional document to help reduce the learning curve. We created first-name only accounts for each researcher so that they were not fully anonymous, but also not carrying any potential stigma associated with status that might be indicated by last names. As they submitted research designs prior to the deliberation, we coded these research designs to develop the most common decisions and potential points of contention. Using this coding we set up three different ‘Kialos’. The deliberations are publicly available, see footnotes.

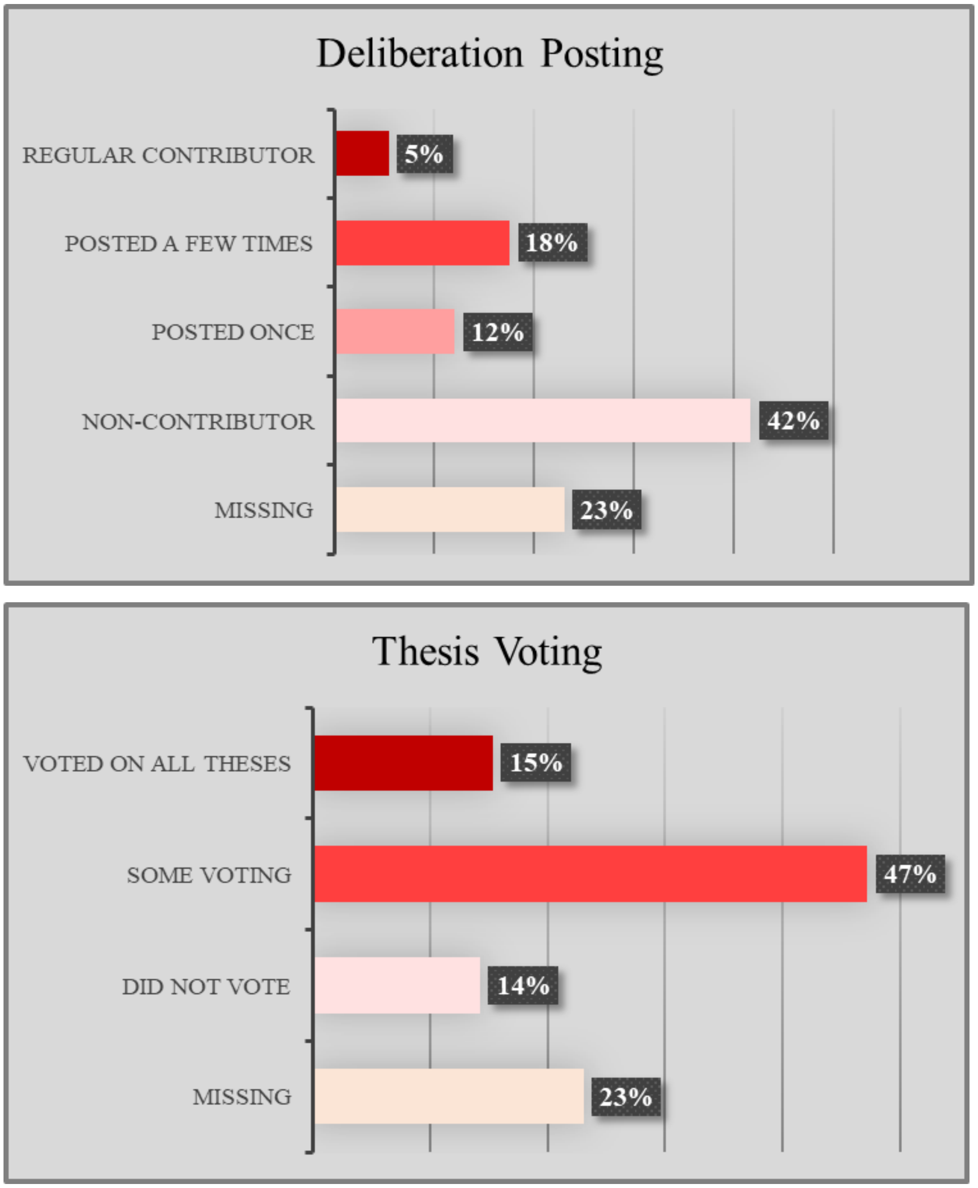
1. Modeling & Estimation[11]
2. Independent Variables & Causal Paths[12]
3. Measurement of Social Policy Preferences[13]

We moderated the process to ensure a safe and respectful deliberation. Also, this allowed us to combine redundant arguments or ask for substance if postings did not have rational arguments in them.

The participation in the deliberation was moderate. In total, of the 91 active CRI participants randomly assigned to the deliberation group, 53% voted on the veracity (‘truthfulness’) of some or all of the deliberation theses presented by the PIs or other participants and only 35% contributed arguments or theses themselves. Figure 4 displays the participation rates as self-reported among the deliberation group. Based on open ended follow up questions, many researchers found it too much to ask that they learn not use a new platform and spend time deliberating on it.

[11] https://www.kialo.com/crowdsourced-replication-initiative---modeling--estimation-19382/
[12] https://www.kialo.com/crowdsourced-replication-initiative---independent-variables--causal-paths-19381/
[13] https://www.kialo.com/crowdsourced-replication-initiative---measurement-of-social-policy-preferences-19380/

Figure 4. Self-Reported Participation in Online Deliberation



Using the results and the qualitative impressions gained through observation of the process provide insights into (1) *if* the researchers made use of the opportunity for deliberation provided to them and (2) *how* the researchers made use of it. We will analyze the survey self-reports on the Kialo experience in the deliberation group. Qualitative content analysis methods will reconstruct frames of knowledge grounded in the observational data. For example, these may indicate positive and negative experiences, contents and forms of knowledge learned and/or shared, differences in the attention and cognitive effort exerted, and where the researchers saw value in the process. Of course, as with any grounded research, what we will find will emerge from the data and could not be known in advance. We expect that the deliberation experience will have a subjective impact on the research participants. We base this expectation on previous research on the consequences of professionally organized and facilitated deliberation events (Davies and Chandler 2012; Escobar 2011; Rose and Sæbø 2010).

In calculating treatment effects on individual researchers, we adjust for the clustered structure of the data using multilevel regression modeling with research teams as a nesting variable. We include the following individual-level covariates from survey Wave 1 to increase the efficiency of the treatment effect estimation: academic experience with multilevel regression, statistics, immigration (all dichotomous), teaching statistics, familiarity with MLM and belief certainty (continuous).

2.5 Phase Three – Expansion

The expansion phase required teams to think about the B&F models and determine if they thought of improvements or alternatives. The idea was to challenge researchers to think about the data-generating model, or plausible alternatives – a crucial goal of replication for robustness testing (Freese and Peterson 2017). This stage started with instructions to write up a research design prior to running any models. Then researchers had the opportunity to revise their designs whether they participated in the deliberation or not. Of those that were invited to participate in the deliberation, 70% of revised their designs after the deliberation. We asked teams to follow their research designs as they planned without deviating. This was another step we took to try and root out biases. Upon finding results they might be unhappy with, we did not want researchers to start searching (i.e., ‘p-hacking’) out different models.

For the expansion we asked each team to report:

> The marginal effect of a 1% higher or lower stock of immigrants, and the marginal effect of a 1 more person per 1,000 (a 1-point increase in net migration) on the dependent variable(s). We ask that you provide 95% confidence intervals for these margins. We realize this may not be possible for all forms of analyses, but please do the best you can to obtain these estimates.

In addition, we asked each team for a subjective conclusion:

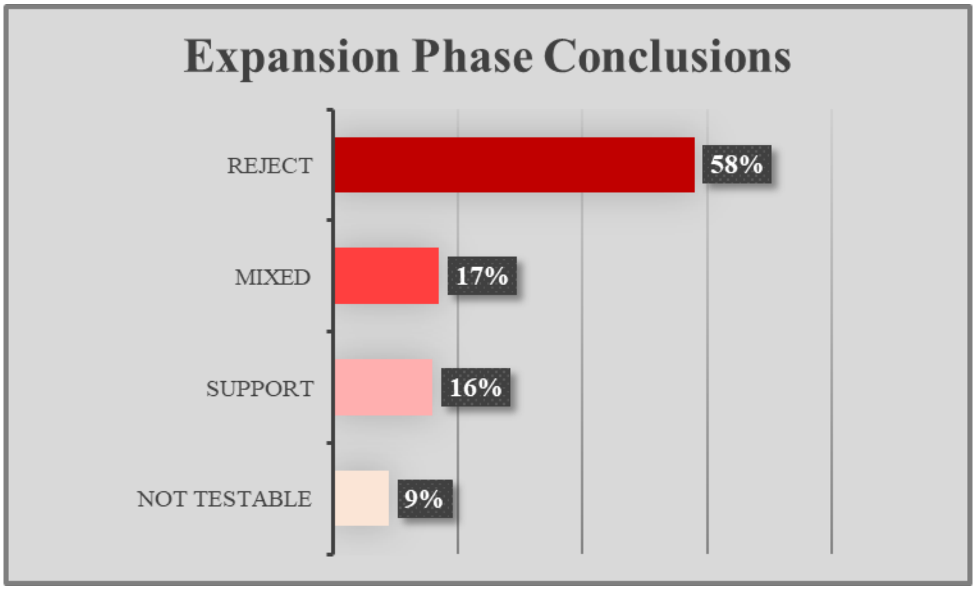
> We ask that you provide a substantive conclusion based on your test of the hypothesis that a greater stock or a greater increase in the stock of foreign persons in a given society leads the general public to become less supportive of social policy, where “social policy” refers to any policy that provides basic protections, social insurance, welfare or wellbeing services, income replacement or active labor market programs. In short, what many scholars refer to as the ‘social welfare state’. Your conclusion should be one of the following options: (a) support, (b) lack of support, or (c) not testable. Importantly, please also provide a short argument (e.g., at least a paragraph) for why you found (a), (b) or (c) as your result[14].

[14] See ‘Communication 8’ https://osf.io/h95kp/

One of our primary interests is to model the results of each team, as in the effect sizes. Using specification curve analyses or what some understand as curation, we plan to determine key variables or research designs that may account for the variance in findings (Orben and Przybylski 2019; Rohrer, Egloff, and Schmukle 2017). Of course we seek to explain subjective conclusions as well; but as a dichotomous outcome we have less variance to explain. Moreover, different researchers might look at the same results and come to different conclusions. This is problematic when trying to perform objective meta-analyses; however, it is very telling about the reality of research. Researchers variability in subjective conclusions based on empirical data is something to consider when looking at the arguments in any given area. For example, the work of B&F has several significant coefficients that support the hypothesis that immigration undermines public preferences for social policy, yet they conclude that is does not.

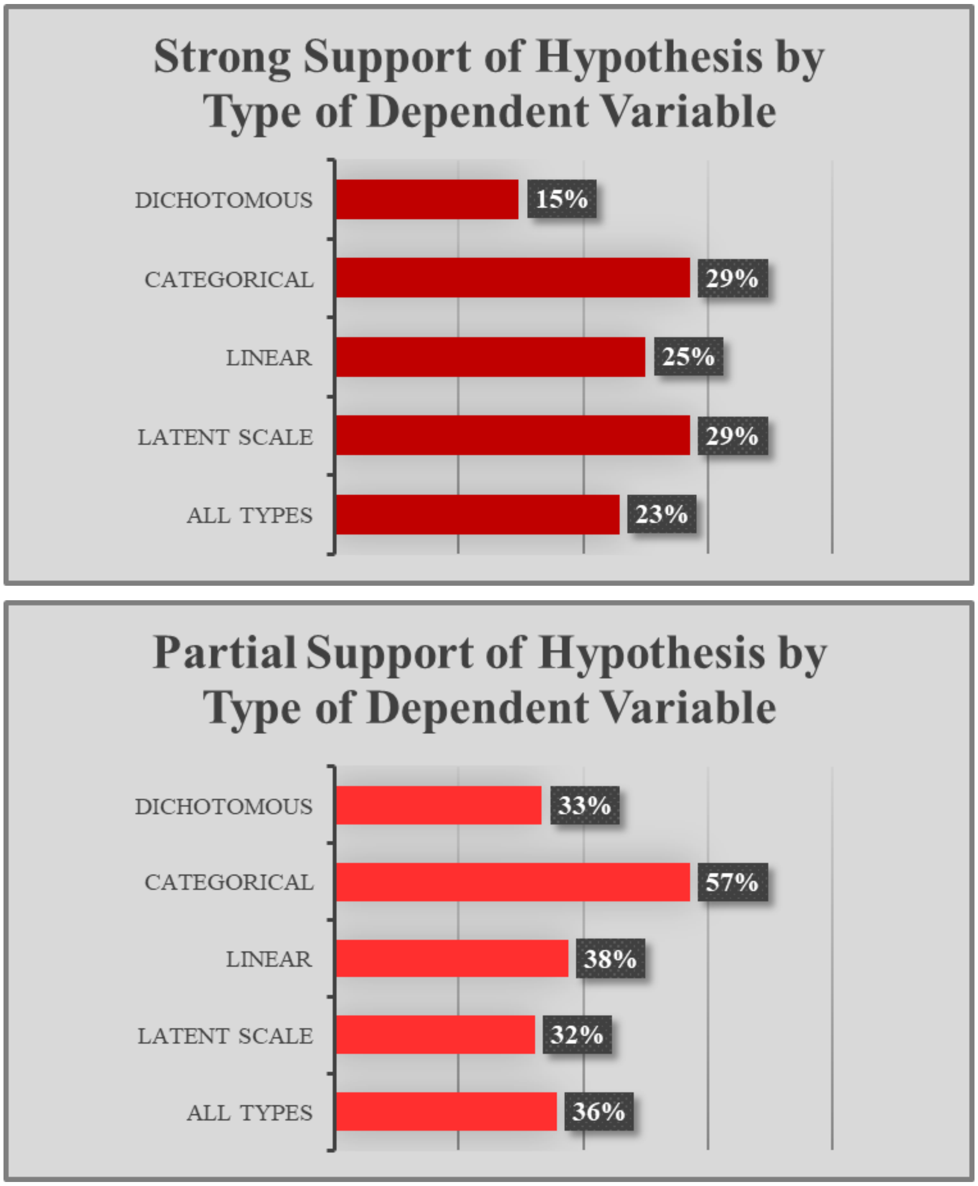
Figure 5 offers the subjective conclusions of the researchers drawn based on their expansion analyses. Figure 6 provides instances of partial support (upper panel) and strong support (lower panel) of the hypothesis that 'immigration undermines public support of social policy' broken down by treatment of the dependent variable.

Figure 5. Results of the Expansion Phase



NOTE: "Reject" indicates support of Brady & Finnigan's (2014) original conclusions

Figure 6. Incidence of Support for the Hypothesis that ‘Immigration Undermines Public Support for Social Policy’



Strong Support of Hypothesis by Type of Dependent Variable

Partial Support of Hypothesis by Type of Dependent Variable

NOTE: “Support” indicates a rejection of Brady & Finnigan’s (2014) original findings

2.6 Phase Four – Post-Result Deliberation & Peer Review

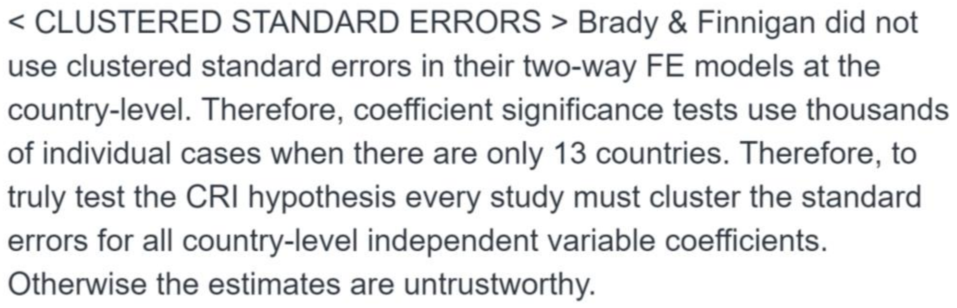
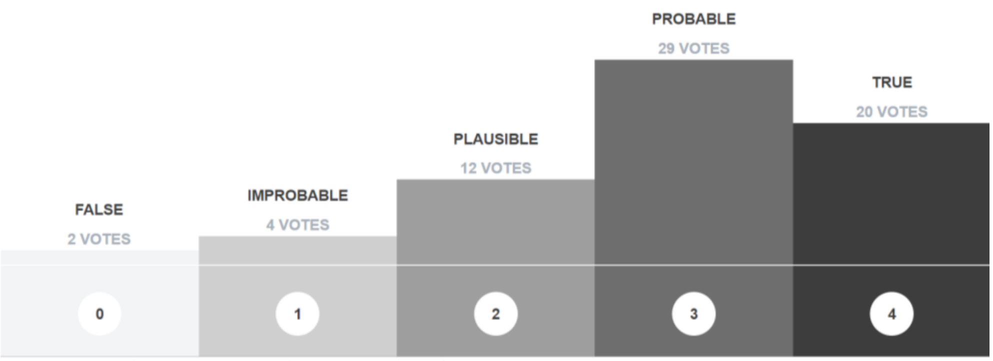
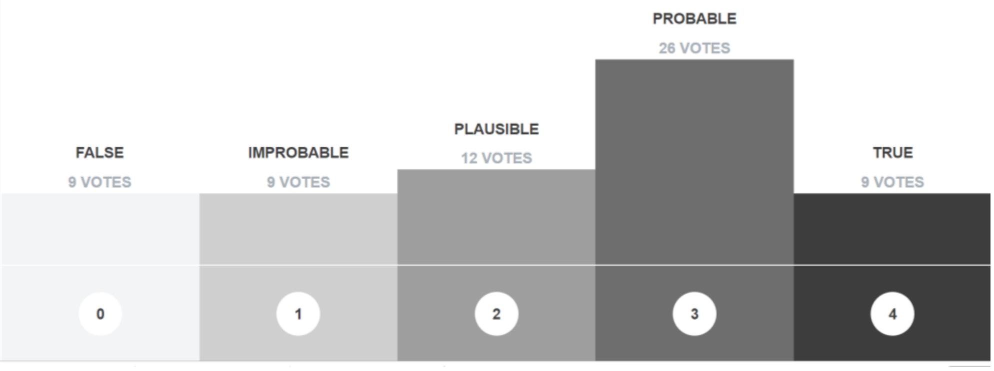
Although we collected research designs, we needed to code the actual research designs in practice from each team based on the code they submitted. This is an area where we see strong improvement for future crowdsourced studies. The research design we received were not standardized. Some teams reported how they measured the dependent variables others not, some teams reported their estimation method, others not. In the future we recommend a standardized research design questionnaire asking for all the specifications of the model. This would also save time coding which took the PIs several weeks of work. However, one problem this would not solve is consistency. Some teams reported model features such as clustering standard errors, or sampling decisions such as including certain waves when they in fact did not incorporate them in their code.

Our original CRI plans were to code and then organize the research designs into prototypical types and then to run a structured deliberation on Kialo with these research designs as the object of discussion. However, the aforementioned delays in review each team's code and the fact that not two research designs were alike, nor did they fall into 'clean' typological categories led us to change course before the post-results deliberation[15]. We extracted the main points of contention from the previous deliberation experiment. We also learned that we gave the participants far too much to deliberate in the first deliberation, therefore in this phase we presented them with just three theses. Below in Figure 7 are the voting results taken directly from the Kialo platform for each by deliberation and control group following the deliberation experiment in phase two.

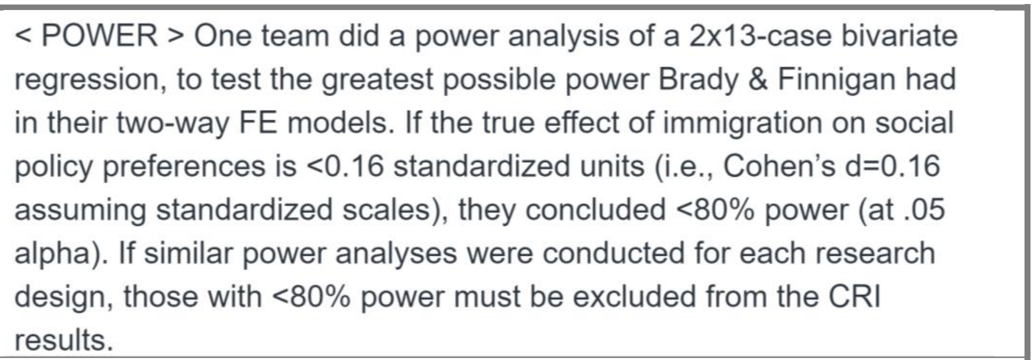
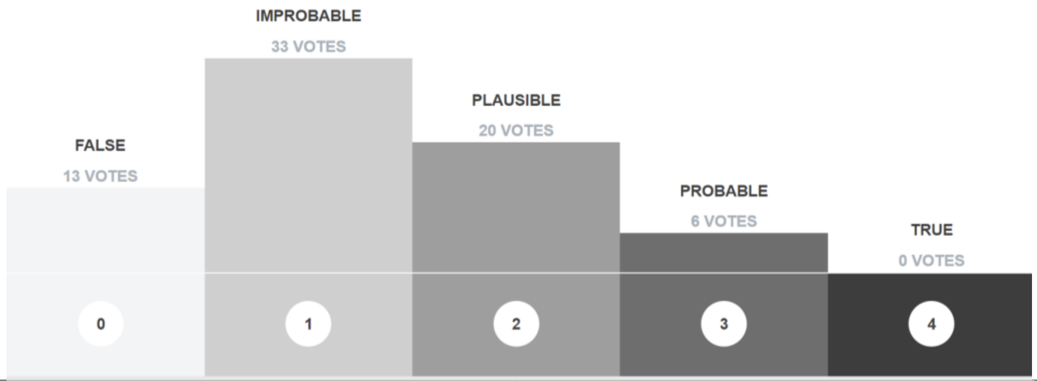
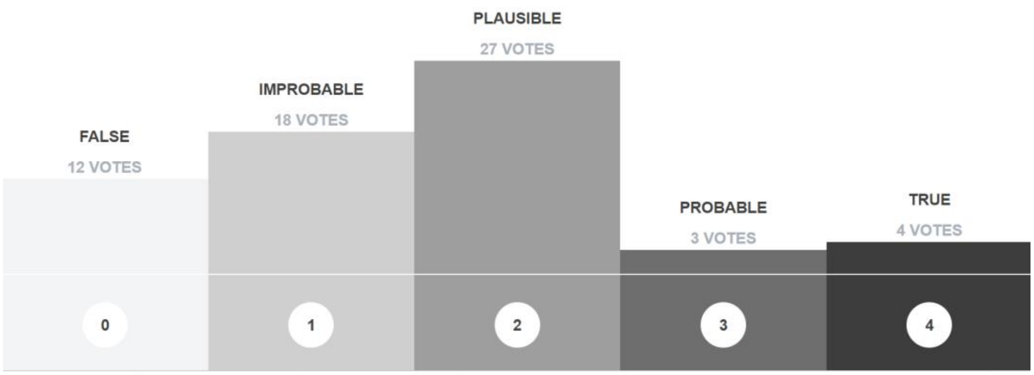
[15] For a list of our coded research designs see Wave 4 or here https://osf.io/9y463/

Figure 7. Kialo Voting Results for Three Critical Research Design Components by Group

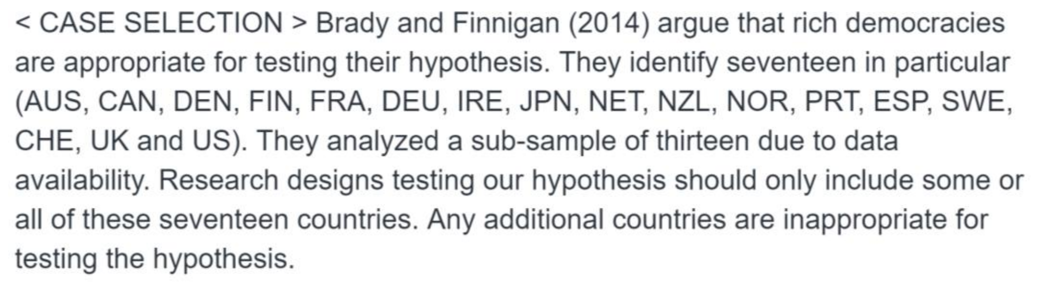
Panel 7A. Clustered Standard Errors

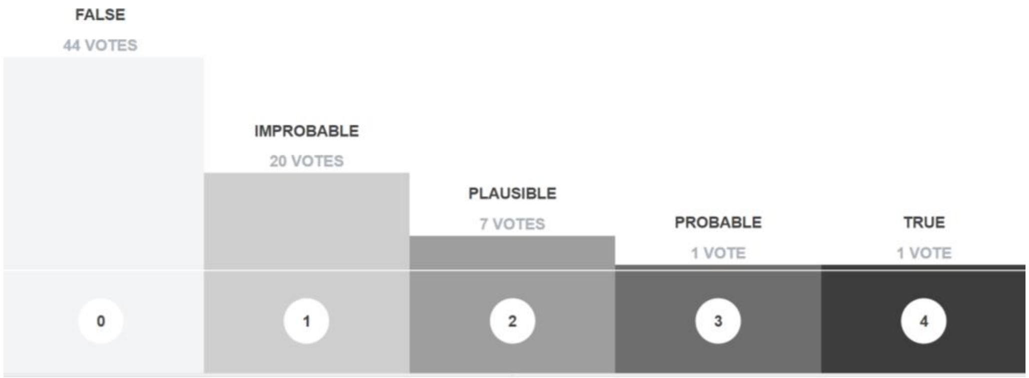
<table>
<tr><td>Thesis</td><td>
&lt; CLUSTERED STANDARD ERRORS &gt; Brady &amp; Finnigan did not use clustered standard errors in their two-way FE models at the country-level. Therefore, coefficient significance tests use thousands of individual cases when there are only 13 countries. Therefore, to truly test the CRI hypothesis every study must cluster the standard errors for all country-level independent variable coefficients. Otherwise the estimates are untrustworthy.</td></tr>
<tr><td>Control Group
<strong>mean = 2.46</strong></td><td>
</td></tr>
<tr><td>Deliberation Group
<strong>mean = 1.80</strong></td><td>
</td></tr>
</table>

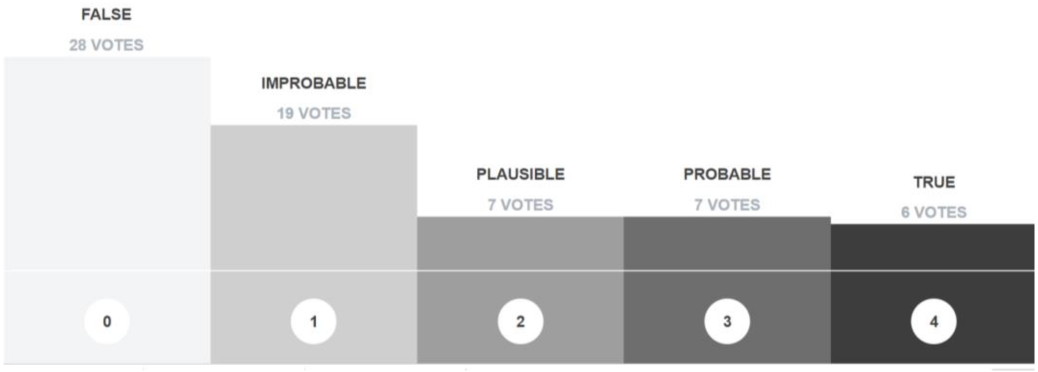
Panel 7B. Power

Thesis	 < POWER > One team did a power analysis of a 2x13-case bivariate regression, to test the greatest possible power Brady & Finnigan had in their two-way FE models. If the true effect of immigration on social policy preferences is <0.16 standardized units (i.e., Cohen's d=0.16 assuming standardized scales), they concluded <80% power (at .05 alpha). If similar power analyses were conducted for each research design, those with <80% power must be excluded from the CRI results.
Control Group **mean = 1.26**	
Deliberation Group **mean = 1.31**	

Panel 7C. Case Selection

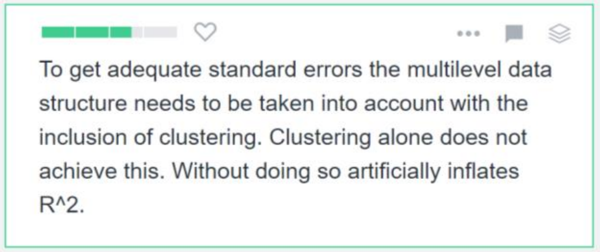
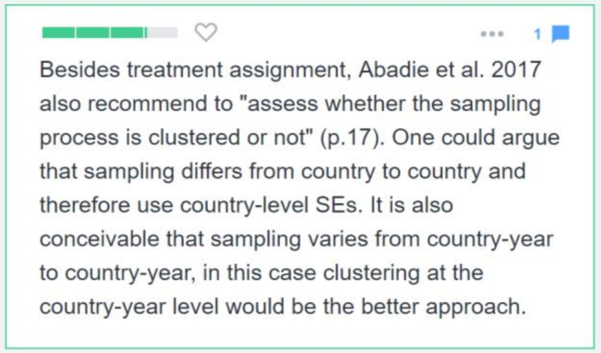
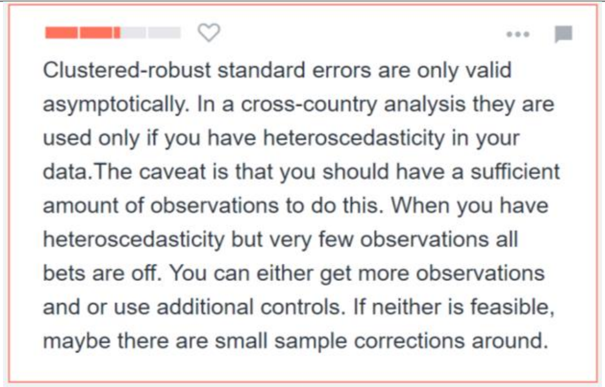
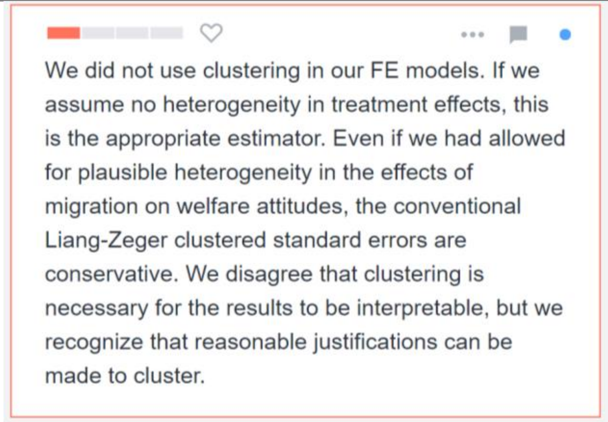
<table>
<tr><td>Thesis</td><td>
&lt; CASE SELECTION &gt; Brady and Finnigan (2014) argue that rich democracies are appropriate for testing their hypothesis. They identify seventeen in particular (AUS, CAN, DEN, FIN, FRA, DEU, IRE, JPN, NET, NZL, NOR, PRT, ESP, SWE, CHE, UK and US). They analyzed a sub-sample of thirteen due to data availability. Research designs testing our hypothesis should only include some or all of these seventeen countries. Any additional countries are inappropriate for testing the hypothesis.</td></tr>
<tr><td>Control Group

<b>mean = 0.56</b></td><td>
</td></tr>
<tr><td>Deliberation Group

<b>mean = 1.16</b></td><td>
</td></tr>
</table>

There is variation between the groups suggesting that the previous deliberation in Phase Two may have influenced participants voting in this phase. What we consider more likely is that the deliberation took on a character of its own in each group and that collectively the voting moved in certain directions, not unlike the public sphere where exposure to particular message persuades preferences and voting behavior (Broockman and Butler 2017; Moy and Rinke 2012). Figure 8 lists potential examples of such messages. A reminder to the reader that there is no difference between the deliberations set up in Kialo for the control or deliberation groups in this phase. We simply keep them separated in their same groups from the Phase Two deliberation experiment to avoid contamination.

Figure 8. Selected Pro and Con Arguments Posted in the Deliberation over Clustered Standard Errors

Control Group Total number of arguments posted = 13 Total votes = 74 Mean = 2.46	Deliberation Group Total number of arguments posted = 5 Total votes = 72 Mean = 1.80
 To get adequate standard errors the multilevel data structure needs to be taken into account with the inclusion of clustering. Clustering alone does not achieve this. Without doing so artificially inflates R^2.	 Besides treatment assignment, Abadie et al. 2017 also recommend to "assess whether the sampling process is clustered or not" (p.17). One could argue that sampling differs from country to country and therefore use country-level SEs. It is also conceivable that sampling varies from country-year to country-year, in this case clustering at the country-year level would be the better approach.
 Clustered-robust standard errors are only valid asymptotically. In a cross-country analysis they are used only if you have heteroscedasticity in your data.The caveat is that you should have a sufficient amount of observations to do this. When you have heteroscedasticity but very few observations all bets are off. You can either get more observations and or use additional controls. If neither is feasible, maybe there are small sample corrections around.	 We did not use clustering in our FE models. If we assume no heterogeneity in treatment effects, this is the appropriate estimator. Even if we had allowed for plausible heterogeneity in the effects of migration on welfare attitudes, the conventional Liang-Zeger clustered standard errors are conservative. We disagree that clustering is necessary for the results to be interpretable, but we recognize that reasonable justifications can be made to cluster.

NOTE: Pro arguments in green borders, Con in red.

3 CONCLUSION

Our main conclusions will arrive in the three working papers we outlined in section 1.3. Here we provided only an overview. We can say with some certainty that the original study of Brady and Finnigan (2014) is verifiable in our replications. The expansions also suggest that their conclusions are robust to a multiverse of data set up and range of alternative model specifications. This suggests there is not a ‘big picture’ finding that immigration erodes popular support for social policy or the welfare state as a whole. However, our findings cast enough suspicion into the equation that further scrutiny is necessary at this big picture level. In particular, our main paper will address some of this necessary further scrutiny by engaging in specification analyses with the intention of explaining variance and model fit in results and seeking out key specifications, data selection or variables that might offer insights into the data-generating models and direct further research in the area.

4 REFERENCES

Alt, James and Torben Iversen. 2017. “Inequality, Labor Market Segmentation, and Preferences for Redistribution.” *American Journal of Political Science* 61(1):21–36.

Baker, Monya. 2016. “Is There a Reproducibility Crisis? A Nature Survey Lifts the Lid on How Researchers View the’crisis Rocking Science and What They Think Will Help.” *Nature* 533(7604):452–55.

Bjerre, Liv, Friederike Römer, and Malisa Zobel. 2019. “The Sensitivity of Country Ranks to Index Construction and Aggregation Choice: The Case of Immigration Policy.” *Policy Studies Journal* 0(0).

Brady, David and Ryan Finnigan. 2014. “Does Immigration Undermine Public Support for Social Policy?” *American Sociological Review* 79(1):17–42.

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.

Breznau, Nate. 2016. “Secondary Observer Effects: Idiosyncratic Errors in Small-N Secondary Data Analysis.” *International Journal of Social Research Methodology* 19(3):301–18.

Breznau, Nate. 2018. “Anti-Immigrant Parties and Western European Society: Analyzing the Role of Immigration and Forecasting Voting.” Working Paper. https://osf.io/8hyrx/

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.

Breznau, Nate, Eike Mark Rinke, and Alexander Wuttke. 2018. “Pre-Registered Analysis Plan for ‘How Reliable Are Replications? Measuring Routine Researcher Variability in Macro-Comparative Secondary Data Analyses’”. https://osf.io/sfuq3

Broockman, David E. and Daniel M. Butler. 2017. “The Causal Effects of Elite Position-Taking on Voter Attitudes: Field Experiments with Elite Communication.” *American Journal of Political Science* 61(1):208–21.

Burgoon, Brian. 2014. “Immigration, Integration, and Support for Redistribution in Europe.” *World Politics* 66(3):365–405.

Clemens, Michael A. 2015. “The Meaning of Failed Replications: A Review and Proposal.” *Journal of Economic Surveys* 31(1):326–42.

Delfanti, Alessandro. 2010. “Open Science, A Complex Movement.” *Journal of Science Communication* 9(3).

Eger, Maureen A. 2010. “Even in Sweden: The Effect of Immigration on Support for Welfare State

Spending." *European Sociological Review* 26(2):203–17.

Eubank, Nicholas. 2016. "Lessons from a Decade of Replications at the Quarterly Journal of Political Science." *PS: Political Science &amp; Politics* 49(2):273–76.

Freese, Jeremy and David Peterson. 2017. "Replication in Social Science." *Annual Review of Sociology* 43(1):147–65.

Gelman, Andrew and Eric Loken. 2014. "The Statistical Crisis in Science." *American Scientist* 102(6):460.

Green-Pedersen, Christoffer. 2004. "The Dependent Variable Problem within the Study of Welfare State Retrenchment: Defining the Problem and Looking for Solutions." *Journal of Comparative Policy Analysis: Research and Practice* 6(1):3–14.

Hjerm, Mikael. 1998. "National Identities, National Pride and Xenophobia: A Comparison of Four Western Countries." *Acta Sociologica* 41(4):335–47.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2(8):e124.

Kesler, Christel and Irene Bloemraad. 2010. "Does Immigration Erode Social Capital? The Conditional Effects of Immigration-Generated Diversity on Trust, Membership, and Participation across 19 Countries, 1981–2000." *Canadian Journal of Political Science* 43(2):319–47.

Kunißen, Katharina. 2018. "From Dependent to Independent Variable: A Critical Assessment of Operationalisations of 'Welfare Stateness' as Macro-Level Indicators in Multilevel Analyses." *Social Indicators Research*.

McLaren, Lauren. 2017. "Immigration, National Identity and Political Trust in European Democracies." *Journal of Ethnic and Migration Studies* 43(3):379–99.

Mitchell, Jeffrey. 2018. "Prejudice in the Classroom: A Longitudinal Analysis of Anti-Immigrant Attitudes." *Ethnic and Racial Studies* 1–20.

Moy, Patricia and Eike Mark Rinke. 2012. "Attitudinal and Behavioral Consequences of Published Opinion Polls." Pp. 225–45 in *Opinion Polls and the Media*, edited by H.-B. C. and S. J. London: Palgrave Macmillan.

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* 0081175018777988.

Orben, Amy and Andrew K. Przybylski. 2019. "The Association between Adolescent Well-Being and

Digital Technology Use." *Nature Human Behaviour*.

Pearl, Judea. 2010. "The Foundations of Causal Inference." *Sociological Methodology* 40(1):75–149.

Pickett, Kate and Richard Wilkinson. 2015. "Income Inequality and Health: A Causal Review." *Social Science & Medicine* 128:316–26.

Rohrer, Julia M., Boris Egloff, and Stefan C. Schmukle. 2017. "Probing Birth-Order Effects on Narrow Traits Using Specification-Curve Analysis." *Psychological Science* 28(12):1821–32.

Saltelli, Andrea and Silvio Funtowicz. 2017. "What Is Science's Crisis Really about?" *Futures* 91:5–11.

Schmidt-Catran, Alexander W. and Dennis C. Spies. 2016. "Immigration and Welfare Support in Germany." *American Sociological Review* .

Schneider, Silke L. 2008. "Anti-Immigrant Attitudes in Europe: Outgroup Size and Perceived Ethnic Threat." *European Sociological Review* 24(1):53–67.

Semyonov, Moshe, Rebeca Raijman, and Anastasia Gorodzeisky. 2006. "The Rise of Anti-Foreigner Sentiment in European Societies, 1988-2000." *American Sociological Review* 71(3):426–49.

Silberzahn, R., E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, F. Bai, C. Bannard, E. Bonnier, R. Carlsson, F. Cheung, G. Christensen, R. Clay, M. A. Craig, A. Dalla Rosa, L. Dam, M. H. Evans, I. Flores Cervantes, N. Fong, M. Gamez-Djokic, A. Glenz, S. Gordon-McKeon, T. J. Heaton, K. Hederos, M. Heene, A. J. Hofelich Mohr, F. Högden, K. Hui, M. Johannesson, J. Kalodimos, E. Kaszubowski, D. M. Kennedy, R. Lei, T. A. Lindsay, S. Liverani, C. R. Madan, D. Molden, E. Molleman, R. D. Morey, L. B. Mulder, B. R. Nijstad, N. G. Pope, B. Pope, J. M. Prenoveau, F. Rink, E. Robusto, H. Roderique, A. Sandberg, E. Schlüter, F. D. Schönbrodt, M. F. Sherman, S. A. Sommer, K. Sotak, S. Spain, C. Spörlein, T. Stafford, L. Stefanutti, S. Tauber, J. Ullrich, M. Vianello, E. J. Wagenmakers, M. Witkowiak, S. Yoon, and B. A. Nosek. 2018. "Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results." *Advances in Methods and Practices in Psychological Science* 1(3):337–56.

Silberzahn, Rafael. 2015. *Crowdsourcing Data Analysis: Do Soccer Referees Give More Red Cards to Dark Skinned Players*. Open Science Framework.

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. "Increasing Transparency Through a Multiverse Analysis." *Perspectives on Psychological Science* 11(5):702–12.

Therborn, Göran. 2013. *The Killing Fields of Inequality*. Cambridge: Polity Press.

Wuttke, Alexander. 2018. "Why Too Many Political Science Findings Cannot Be Trusted and What We Can Do About It: A Review of Meta-Scientific Research and a Call for Academic Reform." *Politische Vierteljahresschrift*.

Wuttke, Alexander, Eike Mark Rinke, and Nate Breznau. 2018. "Deliberative Research: Can Reasoned

Debate Improve the Scientific Process. Registered Research Plan." *OSF Registered Pre-Analysis Plan*. https://osf.io/m2f4a/

Young, Cristobal. 2018. "Model Uncertainty and the Crisis in Science." *Socius* 4:2.