How Many Replicators Does It Take to Achieve Reliability? Investigating Researcher Variability in a Crowdsourced Replication

Author:
Nate Breznau, University of Bremen breznau.nate@gmail.com
0000-0003-4983-3137

Principal investigators:
Nate Breznau, Eike Mark Rinke and Alexander Wuttke

Data assistance:
Hung H.V. Nguyen

Participant replicators:
Muna Adem, Jule Adriaans, Esra Akdeniz, Amalia Alvarez-Benjumea, Henrik Andersen, Daniel Auer, Flavio Azevedo, Oke Bahnsen, Ling Bai, Dave Balzer, Paul C. 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, Eline de Rooij, Alejandro Ecker, Achim Edelmann, Christine Eder, Maureen A. Eger, Simon Ellerbrock, Anna Forke, Andrea Forster, Danilo Freire, Chris Gaasendam, Konstantin Gavras, Vernon Gayle, Theresa Gessler, Timo Gnambs, Amélie Godefroidt, Max Grömping, Martin Groß, Stefan Gruber, Tobias Gummer, Andreas Hadjar, Verena Halbherr, 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, Sabine Israel, Laura Jacobs, Jannes Jacobsen, Bastian Jaeger, Sebastian Jungkunz, Nils Jungmann, Jennifer Kanjana, Mathias Kauff, Sayak KhatuaManuel Kleinert, Julia Klinger, Jan-Philipp Kolb, Marta Kołczyńska, John Kuk, Katharina Kunißen, Jennifer Kanjana, Salman Khan, Dafina Kurti, Alexander Langenkamp, Robin Lee, David Liu, 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, Robert McDonnell, Patricia McManus, Kyle McWagner, 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, Mark Pickup, Christopher Prosser, Louis Raes, Kevin Ralston, Miguel Ramos, Frank Reichert, Leticia Rettore Micheli, Arne Roets, Jonathan Rogers, Guido Ropers, Robin Samuel, Gregor 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, Jordan W Suchow, Janna Teltemann, Andrey Tibajev, Brian Tung, Giacomo Vagni, Jasper Van Assche, Meta van der Linden, Jolanda van der Noll, Arno Van Hootegem, Stefan Vogtenhuber, Bogdan Voicu, Fieke Wagemans, Nadja Wehl, Hannah Werner, Brenton Wiernik, Fabian Winter, Christof Wolf, Cary Wu, Yuki Yamada, Nan Zhang, Conrad Ziller, Björn Zakula, Stefan Zins, Tomasz Żółtak

ABSTRACT

The paper reports findings from a crowdsourced replication. Eighty-four replicator teams attempted to verify results reported in an original study by running the same models with the same data. The replication involved an experimental condition. A “transparent” group received the original study and code, and an “opaque” group received the same underlying study but with only a methods section and description of the regression coefficients without size or significance, and no code. The transparent group mostly verified the original study (95.5%), while the opaque group had less success (89.4%). Qualitative investigation of the replicators’ workflows reveals many causes of non-verification. Two categories of these causes are hypothesized, routine and non-routine. After correcting non-routine errors in the research process to ensure that the results reflect a level of quality that should be present in ‘real-world’ research, the rate of verification was 96.1% in the transparent group and 92.4% in the opaque group. Two conclusions follow: (1) Although high, the verification rate suggests that it would take a minimum of three replicators per study to achieve replication reliability of at least 95% confidence assuming ecological validity in this controlled setting, and (2) like any type of scientific research, replication is prone to errors that derive from routine and undeliberate actions in the research process. The latter suggests that idiosyncratic researcher variability might provide a key to understanding part of the “reliability crisis” in social and behavioral science and is a reminder of the importance of transparent and well documented workflows.

1 RESEARCHER VARIABILITY[1]

Any given study confronts researchers with various decisions and potential actions to take in design, measurement, analysis and result reporting (Wicherts et al. 2016). Any researcher might use its universe of potential choices differently leading to variation in results and conclusions across researchers (Gelman and Loken 2014). Perhaps this is one reason many social and behavioral researchers failed to replicate previous findings (Maxwell, Lau, and Howard 2015; Open Science Collaboration 2015). The problem is particularly acute in experimental research where new samples are drawn and instruments may not be identical or identically implemented as the original; however, in non-experimental involving repeated use of publicly available secondary data the problem may also exist (Thompson et al. 2020; Young 2018). I refer to this as *researcher variability*, defined as different outcomes among researchers ostensibly testing the same hypothesis with the same data. I suggest it is another form of potential noise worthy of social scientists' consideration.

To investigate this phenomenon, I look at data collected in a crowdsourced study involving the simplest form of replication: a *verification* (Freese and Peterson 2017)[2] defined by an act of checking if the original data and reported models reproduce the reported results. This provides a most conservative test case, because researchers have few if any decisions to make and reliability should be very high. None the less, I expect even verifications are prone to uncertainty. For example, the *American Journal of Political Science* introduced external checking of code and results and it took an average of 1.7 re-submissions of the code per article before results were verifiable (Jacoby, Lafferty-Hess, and Christian 2017; Janz 2015). In a similar vein, Hardwicke et al. (2018) attempted to replicate studies published in the journal *Cognition* and found that even with author assistance, 37% of the articles (13-of-35) had at least one effect that could not be statistically reproduced within 10% of the original. Stockemer et al. (2018) showed that among major opinion and electoral research publications in 2015, one-third could not be verified and one-quarter could not produce any results because the code was so poorly organized. These and other studies suggest that even basic 'copy-paste' research is prone to variability (see also Eubank 2016).

Whether researcher variability exists is somewhat unknowable because it is by definition a meta form of uncertainty. The fundamental research question is whether different researchers, in a parallel

[1] As I, Nate Breznau, am author and analyst of the results presented in this paper it is written in the first person. Comments were provided by many participants and the original principal investigators, and all are included as co-authors as their research efforts in this project warrants. I opt for "it" as opposed to "he" or "she" as the preferred replicator pronoun throughout this article.

[2] I use "verification" here following Freese and Peterson (2017), also sometimes known as a "reproduction"; i.e., verifying that the reported models reproduce the reported results. This is not verification in a philosophical or Popperian sense, whereby it is arguable that nothing can unambiguously be "verified" or "true".

universe where all other conditions are identical, would sometimes come to different results. This is not currently testable; however, the data analyzed herein offers some insights based on a crowdsourcing design with controlled conditions.

1. The Threat of Researcher Variability

I propose that researcher variability is a problem if it leads to an error rate of 5% or more. In other words, if 95% of a population of potential replicators would verify the original results if they engaged in a replication then one replication should be sufficient to trust the results. If the verifiability of a study could magically be known in advance, then the error rate would simply be the rate at which researchers are unable to verify the results relative to this prior verifiability rate. But the true prior rate of verifiability is difficult if not impossible to observe. A failed verification, like any failed replication, could indicate an unverifiable study, or that the replicator made an error. This makes the present study ideal for an exploratory effort to understand researcher variability, because the prior probability of the original results should be equal to 1.00 (i.e., 100%) because it was known in advance that exact same data, models and code from the original study produce the reported results.

2. Two Potential Types of Researcher Variability

The idea of idiosyncratic researcher variability comes from the pre-analysis plan posted by the researchers that collected the data for this study (Breznau, Rinke, and Wuttke 2018). These researchers categorized researcher variability as either “non-routine” – mistakes or intentional variation, or “routine” – undeliberate or idiosyncratic variation. Some examples should help illustrate the difference. One of the crowdsourced replicators insisted that the authors of the original study made a mistake because they did not cluster standard errors at the country-level. This replicator therefore added clustering to its verification analysis and introduced intentional decisions as a form of analytical flexibility. This led to identical effect sizes, but different significance levels. Perhaps more importantly, it changed the model specification so that it was no longer technically a verification but a *reanalysis* (same data, different models) (Christensen, Freese, and Miguel 2019). This is a case of **non-routine** variability. The replicator intentionally did something different that was not part of the research design.

Alternatively, a different replicator submitted results rounded to two decimal places, thus appearing mathematically different from the original three-decimal-place results. This case is **routine** variability because the replicator may or may not have consciously decided to conduct rounding, it could be a product of software defaults for example. Therefore, to use the data from Breznau, Rinke and Wuttke (2018) to investigate researcher variability as representative of real-world research requires discrete treatment of routine researcher variability on the one hand, as resulting from undeliberate actions within constraints, from non-routine researcher variability on the other, as resulting from liminal and deliberate actions including mistakes in achieving some research goal or following a pre-conceived research design. Table 1 is a hypothetical and non-exhaustive list from the aforementioned pre-analysis plan.

Table 1. Distinguishing Two Forms of Researcher Variability in Replications

Source	Routine	Non-Routine
Mistakes	Minor coding or reporting mistakes - idiosyncratic events or traits	Major coding or reporting mistakes; 'sloppy science'
Expertise of the researcher	Unclear – although may reduce variability due to homophily of expert replicators and original authors	Should reduce mistakes as a function of method skills
Modeling	Unintentional features of a model generated in the model construction phase	Deliberate decisions in constructing a formal or statistical model
Software	A researcher's standard software type, packages or version (their personal ‘defaults’) The defaults of the software, packages and version.	Possibly the level of experience with that software, method or package Exception: advanced users might override these defaults
Extra steps	Unintentionally adding or altering an analysis, something not mentioned in an original study 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	Exception (arguably): using illegal channels to gain access to software or data, although I take no ethical stance in this paper
'Random' Error	Variability that cannot be controlled, often undetected	Variability accounted for or explained as intentional research choices or implicit data-generating process claims
Quality / Transparency of materials	Forces researchers to work harder/take more steps, introducing more opportunities for routine error	Forces researchers to make more choices, introducing more opportunities for non-routine error

NOTE: Table from pre-analysis plan of the *Crowdsourced Replication Initiative* and reflects hypothesized causes of researcher variability.

An example of the subtleness of routine variability is that different researchers have different methodological or disciplinary training and institutional access to software leading to modes of operation that are not intentional across researchers. These contexts are a product of what is known and available, thus idiosyncratic to researchers. This type of error is not readily observable, or easily categorized as a mistake or a choice. For example, does lack of access to version 16 of *Stata* constitute a choice or a mistake? I would argue it is neither. Alternatively, consider a scenario where researchers publish a study and their regression models are completely verifiable. The code is shared publicly and replicators should reproduce the exact same results every time. However, replicators might need to download the data from a repository and the data may have changed, something some archives do without version control (Breznau 2016). Even if the data are the same, differences between operating systems, packages and even processors can produce different results (McCoach et al. 2018). Default settings often change across versions or software packages. Again, these problems are not really mistakes or choices, they just happen without much thought. Routine researcher variability is something unconscious, or preconscious, to the process of research and the researcher, something hard to observe and not part of the modeling process or research design. It depends largely on tacit knowledge that researchers require to execute their studies that "cannot be fully explicated or absolutely established" in practice (Collins 1985:73).

Adjudicating between these two types is necessary, because non-routine researcher variability constitutes mistakes or deviations from what should have been the research design. Careful attention to the methods by the researcher, reviewers, replicators and potentially meta-analysts should identify if not eliminate such mistakes; at least in an ideal research world. If I hope to observe routine researcher variability I need to eliminate these non-routine mistakes from the data. Then what remains should constitute routine researcher variability and make it possible to test the hypothesis that no matter how carefully researchers follow a research design and analysis, subtle variation in results will occur across researchers.

2 USING A CROWDSOURCED REPLICATION TO MEASURE RESEARCHER VARIABILITY

In 2018, Breznau, Rinke and Wuttke (2019) launched the *Crowdsourced Replication Initiative* (CRI). This was a study testing both reliability in social science and a sociologically meaningful hypothesis. The first phase of the project was to collect a pool of researchers and observe them as they verified a study's previously published quantitative results. The data from this first phase is the focus of this paper.

The CRI elected to replicate David Brady and Ryan Finnigan's (2014) *American Sociological Review* study titled, "Does Immigration Undermine Public Support for Social Policy?". This original study met several ideal criteria: highly cited, freely available data and code, independently verifiable by two of the CRI's principal investigators and the original authors were comfortable with their work being the target. The Brady and Finnigan study used *International Social Survey Program* (ISSP) data with questions about the government's responsibility to provide various forms of social welfare and security. The study aggregated responses to these questions and regressed them on stock and flow of immigration measures at the country-level across different model specifications including social spending and employment rate as independent variables.

Power analyses determined that at least 60 replicator teams were necessary to introduce an experimental design (Breznau et al. 2018). Fortunately, 105 teams of between one and a maximum of three persons registered and 99 successfully completed the first CRI task which was to respond to a survey. Random assignment of the original 99 replicator teams placed 50 into a *transparent group* that received the Brady and Finnigan published paper, the Stata code and a published technical appendix. This group had minimal research design decisions to make, just verify the original study following their methods and code. The other 49 teams, the *opaque group*, got an anonymized and far less transparent version of the study. It was a derivation of the Brady and Finnigan study altered by the CRI principal investigators. It included only 4-out-of-6 of the dependent variables, an analysis without the individual-level income variable (selected for removal because it had no noticeable impact on any results) and instead of the full paper got only a 'methods section' written by the principal investigators describing the models, direction and significance of coefficients without any numbers and without code (see Appendix 7). This offered an experimental condition simulating polar extremes in the transparency of an original study's materials. For the purposes of simulating a real research endeavor, the participants were instructed to use the software they were most comfortable with, rather than learn Stata. In the transparent group the Stata users were asked to please write their own version of the original code rather than simply run the file from the original authors.

Participants had three weeks to complete the replication, with extensions granted upon request. Participants were asked to present odds-ratios following the original study. All participating replicators received a template to limit errors associated with reporting. Four models included both stock and flow measures of immigration (percent foreign-born and net migration), this meant 12 of the models produced 24 odd-ratios leading to 48 results from 36 models (Brady and Finnigan 2014: Tables 4 & 5). Each team in the transparent group reported 48 odds-ratios and each in the opaque group reported 40 because they did not get the model with both immigration measures at once, another step to hide the identity of the original study. A few models ran into non-convergence problems therefore not all reported all results. The final N was 3,695 odds-ratios from 85 teams.

Multiple CRI participants did not consent to sharing their replication code given that it might show differences in skills and reveal mistakes. Therefore, the codes from this study is not publicly available; however, in the interest of open science, scholars may request permission to view them so long as they sign a confidentiality agreement. Except for the original codes, readers will find all the shareable data, analyses and replication files in the Project Repository[3].

For the quantitative aspect of this research, I coded the types of results: *Verification* (dummy) if the odds-ratio went in the same direction meaning above, below or equal to one (at +/- 0.01), *Exact Verification* (dummy) if odds-ratio was numerically identical to the second decimal place (< 0.01), and *Deviance* (continuous) as the absolute value of the difference of the estimated odds-ratio and original. The distribution of these measures are presented in Table 2. Based on a participant survey, I constructed variables for the discipline of the replicator, taking the majority discipline or first discipline for teams of more than one person. I collapse this into a variable labeled as *Sociology* where sociology degrees = 1 (43 teams) and political science = 0 (22 teams). Other degrees did not have enough cases for meaningful comparison (e.g., psychology, communications, methods-focused or economics). Then I created a variable *Stats-Skill* as a latent factor from 4 questions on their experience with statistics and their subjective skills. I also create a variable from one question called *Difficult* reflecting a score of 0-5 where a 5 indicates that the replication was subjectively the most difficult. I code statistical software as *Stata* = 1 (56 teams) versus other software (22 used R, 4 used SPSS and 3 used Mplus).

For the qualitative aspect of this research, I reviewed the results and replication code of all teams. A research assistant helped ensure that the code could be run prior to my qualitative analysis of the code. In only two cases were further exchanges with the authors necessary to get their code running. Once certain that the code would produce the results submitted by each team, I identified and categorized sources of researcher variability. A semi-directed typology of researcher variability emerged from the ideas in Table 1 plus some new forms. I identified the difference between mistakes or deliberate deviations from the research design (non-routine) and procedural aspects (routine). I then corrected mistakes if it was obvious what the team would have done in a counterfactual scenario. I only changed code when I did not have to make *any* decisions. For example, if a team omitted a ‘fixed-effect’ for country or year, I corrected this. If a team forgot to include a country or added an extra country into the original sample of 13, I adjusted this. However, if I had to make recoding decisions that could have been interpreted in different ways given the description of the research design, like how to standardize or collapse categories of employment or education by country, I did not take any action. It was theoretically possible to eliminate all sources of deviation from the original results in all but one teams’ code; however,

[3] Will be publicly shared, anonymized R markdown files are available in the supplementary materials of this submission.

I used a rule that if I could not identify the source within two hours of reviewing and running the code and if the verification rate was higher than 95%, I would not report any qualitative types of variability. The motivation for this was that real-world research is not perfect and researchers can expect only so much scrutiny from reviewers, editors, other researchers and their own critical reflections.

After eliminating non-routine researcher variability counterfactually wherever possible I had a new curated set of results from which I recalculated the rates of *Verification*, *Exact Verification* and *Deviance*. These should present a more realistic scenario of what would have been discovered in a real-world research process. I also created a new quantitative measure taking on a value of one if I identified any form of *Routine* researcher variability in their code.

3 RESULTS

I first present sample means for all estimated effect sizes in the first three numerical columns of Table 2, under "Means by Sample". These are presented in ratio format, but can be interpreted as percentages in most cases. The "Raw" results are in the first three numeric rows. These reflect exactly what was submitted by the teams. Descriptively, 95.6% of estimated effects in the transparent group (first column) were a *Verification* and 77.2% were an *Exact Verification*. The mean *Deviance* from the original study in this group was 0.014. These statistics drop somewhat in the opaque group with only 89.4% *Verification*, and 48.2% *Exact Verification* and a mean *Deviance* of 0.071. The third column reports the totals for both groups. The next three numeric rows are "Curated" results after adjusting the code for non-routine errors where possible. The transparent group had 96.8% *Verification*, 82.2% *Exact Verification* and 0.011 mean *Deviance*, while the opaque group had 92.3%, 56.6% and 0.017 respectively. The right portion of Table 2 presents correlations that provide the main quantitative findings that I discuss toward the end of the Results section.

Table 2. Means and Correlations of Raw and Curated Replication Outcomes

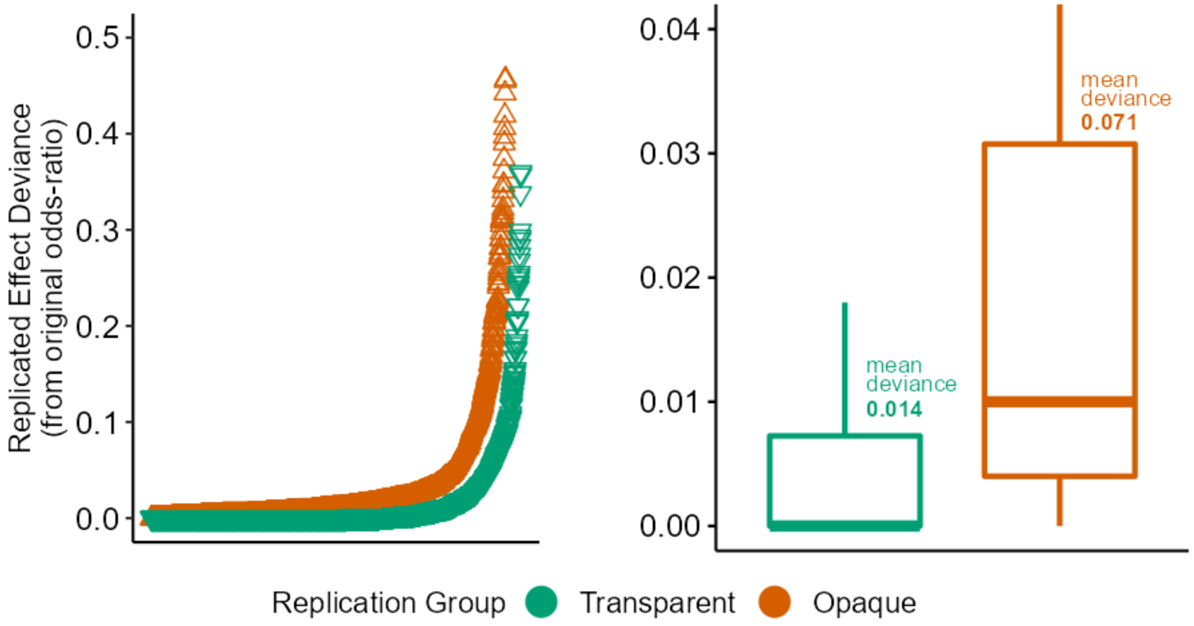
Variables	*Measurement*	*Means by Sample*			*Correlations w/ Raw Results*			*Correlations w/ Curated Results*		
		Trans-parent	Opaque	Pooled	Verif-ication	Exact Verif.	Deviance	Verif-ication	Exact Verif.	Deviance
Raw Replication Results										
Verification	same direction =1	0.956	0.894	0.924	1					
Exact Verification	identical at two decimals =1	0.772	0.482	0.626	0.370	1				
Deviance	absolute diff. w/ original	0.014	0.071	0.043	-0.301	-0.352	1			
Curated Replication Results										
Verification	same direction =1	0.968	0.923	0.946	--	--	--	1		
Exact Verification	identical at two decimals =1	0.822	0.566	0.694	--	--	--	0.366	1	
Deviance	absolute diff. w/ original	0.010	0.025	0.017	--	--	--	-0.613	-0.522	1
Independent Variables[a]										
Stata	other software =0	0.667	0.620	0.644	**0.132**	**0.217**	**-0.108**	**0.138**	**0.185**	**-0.170**
Sociology Degree	political science =0	0.487	0.511	0.499	0.036	-0.021	0.045	0.033	-0.001	**-0.075**
Stats-Skill	continuous scale, standardized	-0.047	0.093	0.023	-0.002	-0.036	**-0.049**	0.006	-0.022	0.000
Difficulty	5-point scale standardized	-0.088	0.002	-0.043	**-0.176**	**-0.226**	-0.008	**-0.224**	**-0.239**	**0.292**
Team Size	1-3 persons	2.103	2.253	2.178	-0.036	-0.004	**0.067**	-0.040	0.004	0.018
Routine	routine variability =1	0.308	0.747	0.527	**-0.081**	**-0.343**	0.031	**-0.074**	**-0.345**	**0.228**
Transparent	transparent group =1	1.000	0.000	0.500	**0.113**	**0.282**	**-0.191**	**0.099**	**0.277**	**-0.171**

NOTE: Effect are odds-ratios. *Verification* are same direction of effects, *Exact Verification* are identical at <0.01 and *Deviance* is absolute value of the difference between estimated odds-ratio and original odds-ratio. Sample here refers to 3,742 estimated effect sizes from 85 replicator teams, 1,872 (39 teams) in the transparent group and 1,870 (46 teams) in the opaque group.

[a] Bold correlations significantly different from zero at p<0.001

To visualize the impact of transparency, Figure 1 plots the absolute deviance of the estimated effect sizes from the originals. The left panel demonstrates that the opaque group's results (orange upward triangles) are on average further from an exact verification than the transparent group (green downward triangles). The right panel demonstrates that the median transparent group replication was 0.000, an exact numerical replication (mean deviance = 0.014), while the opaque group had a median of 0.011 (mean = 0.071) and a wider dispersion (larger box, indicating greater inter-quartile range). This means that in the opaque group less than 50% of the estimated effects were an exact verification. Clearly having less transparent materials makes replication more error prone.

Figure 1. Deviance of Verifications, 3,695 estimated odd-ratios reported by 85 replication teams



Note: Models ordered by Deviance in left panel

Turning to the causes of researcher variability, my qualitative investigation is summarized in Table 3. There were roughly four major categories of causes of variance in variance in the dependent variables. "Major mistakes" and "Lesser mistakes" were rare but impactful factors. All mistakes are non-routine researcher variability by definition. I expect these would be eliminated under normal research conditions if the research was being prepared for an academic journal or other format for public scrutiny. The column "Counterfactual" indicates if it was possible to correct the mistakes in the process of curation.

The third type of variability was "Different analytical processes". Here, the most common are put in bold. Researchers routinely made slightly different coding decisions than the original. For example, many teams recoded employment status of “helping a family member” into “not in labor force” when the original study coded this as “part-time work”. Others coded this same variable as “unemployed” and some coded “unemployed” as “not in labor force”. Two teams disaggregated this variable into “full” and “part-time” based on a third variable measuring hours of work per week. There are many analogous instances with employment status, education level and occasionally income (among the transparent group). An annotated list by team is available in Appendix B. In the case of the transparent group, these recoding decisions were far less common, presumably because the replicators could look at the code and see exactly which categories were recoded into full-time, part-time, unemployed and not in labor force for example. The opaque group did not have this option. Treatment of missing was a perpetual source of variation. Some used listwise on all variables, some on all four dependent variables and others did no deletion of missing letting the software instead remove them for each model. A peculiar problem arose in some cases where dummy variables were coded with the object of interest as “1” (like ‘in labor force’) and then all others (including true missing values) coded as “0” meaning that values were added to the analysis that were dropped in the original study.

Table 3. Observed Researcher Variability among 85 Replicator Teams

Type	Examples	Counter-factual[a]	Researcher Variability	# of Teams
Major mistakes	• Recoded all values on the dependent variable to zero in one wave	No	Non-Routine	1
	• Did not include individual-level control variables	No	Non-Routine	2
	• Recoded all/certain categories of a dependent variable to 0	Yes	Non-Routine	2
	• Regression intercepts larger than 1 trillion (re-coding mistakes)	Unclear	Non-Routine	1
	• Ran separate analyses by wave	No	Non-Routine	2
Lesser mistakes	• Reporting error. Wrong models' odds-ratios or other mismatch submitted in results	Yes	Non-Routine	5
	• Forgot year dummies ('fixed-effects')	Yes	Non-Routine	1
	• Reverse coded 1996 & 2006 as years	Yes	Non-Routine	1
	• Slightly different sample of countries analyzed	Yes	Non-Routine	2
	• Used linear or multilevel-logistic instead of logistic regression	Yes	Non-Routine	3
	• Used one or more extra independent variables	Yes	Non-Routine	2
Different analytical processes	• **Categorical differences coding socio-economic variables**	**No**	**Routine**	**42**
	• **Different treatment of missing (listwise for all DVs, dropping a category from a control variable, recoding missing on income to zero)**	**No**	**Routine**	**8**
	• Used robust clustered SEs	Yes	Non-Routine	5
	• Kept only former West Germany but dropped former East	No	Routine[c]	1
	• Generated results with only two decimal places	No	Routine	2
Context and idiosyncrasies	• Type of software used[b]	No	Routine	NA
	• Version of software or software package used	Maybe	Routine	NA
	• Skills of the researchers	No	Routine	NA
	• Discipline or major area of study	No	Routine	NA
	• The quality of materials provided. Degree of transparency in the case of replication.	No	Routine	NA
	• Intransparent – steps of the coding process missing, e.g., done by point-and-click or not saved	No	Routine	4

Note: "NA" means not applicable because it is not quantifiable and/or theoretically applies to all teams.

[a] Counterfactuals were only used when it was possible to change the specific decision or mistake without changing any other aspects of the code or making any decisions on behalf of the team.

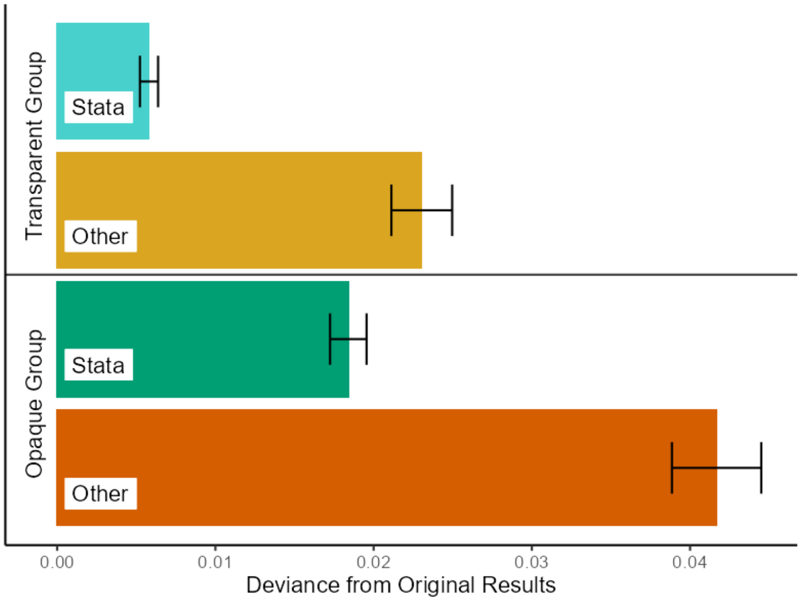
[b] Concretely, one team reported that had different team members recode the datasets in *Stata* and *R* and then compared results to find that they were slightly different.

[c] Debatable. This may reflect a standard practice for those who regularly work with German surveys, or an intentional choice.

The final category in Table 3, “Contexts and idiosyncrasies” comes mostly from the submitted results and the researcher survey rather than the submitted code, for example software type, discipline, statistics-skills and reported difficulty with the replication. The awareness that version of the software might matter became obvious when replicating the teams’ code. Often newer versions of Stata or R would not run due to changes in the language or deprecated file formats. The impact of these routine variabilities is somewhat unknown because there is no possibility to observe the same researcher doing this research using a different software or with a different level of skills. One team did the replication in both Stata and R. They only reported the Stata results because they were exact, but mentioned in an email that the R results were slightly different for “unknown reasons”. Out of necessity I had to get the code to run, thereby changing packages or syntax without making qualitative changes was necessary so long as the changed code produced their reported results.

To help shed further light on “Context and idiosyncracies” I turn back to quantitative analysis using correlations presented in the right portion of Table 2. It is clear that researchers using Stata were more likely to verify (r = 0.132 for *Verification* and 0.217 for *Exact Verification*) and less likely to have *Deviance* (r = -0.108). The bold in Table 2 indicates these are significant Pearson correlations at p<0.001. These associations remained after curating the mistakes. I visualize this phenomenon in Figure 2 using the team scores on the variable *Deviance*. Stata users in the transparent group were far more likely to verify the original with a *Deviance* near zero. This makes sense for the transparent group because Stata users could simply reproduce the code provided to them. Even if they followed the replication instructions to write their own code, they had a recipe to follow. Moreover, Stata users speak the ‘language’ of the original study’s analysis and presumably can understand it much better than non-Stata-users. Although the opaque group (bottom panel) had higher overall *Deviance*, the Stata users within that group had lower *Deviance*, even slightly lower than the *Deviance* of the non-Stata users in the transparent group. This is striking. The opaque group were given no code whatsoever to look at. Moreover, the experimenters asked them after debriefing whether they recognized the original study, and all participants said “no”.

Figure 2. The Role of Same Software in Replication Reliability

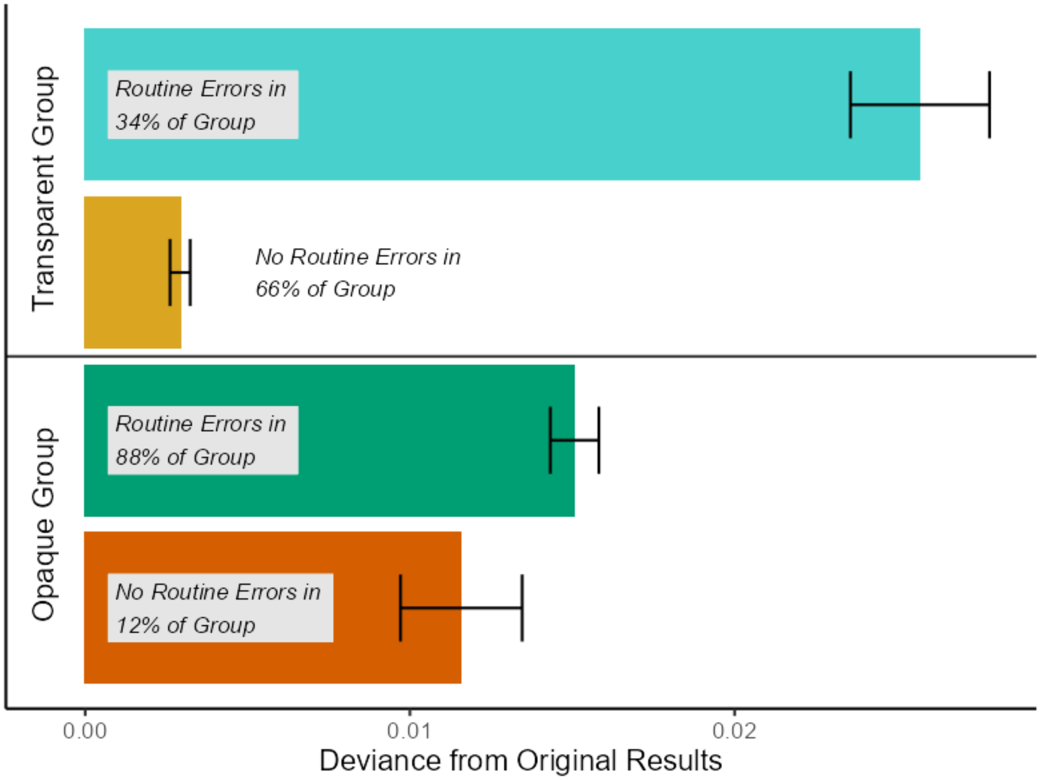


Note: Deviance = absolute difference between reported odds-ratios and the original study. The original study was done with Stata software.

As for the other variables' correlations, there is little difference between sociology and political science degrees. In the Project Repository readers will see that lumping all the other degree types together leads to a category less likely to replicate, but there is no meaning to this category as it has 1-6 researchers from each of psychology, communications, economics and some interdisciplinary degrees. Inference from such tiny samples seems unwise. There are no differences by statistics-skills, at least not after curation, contrary to the hypotheses of the CRI principals (see Table 1). Those who reported a higher degree of difficulty conducting the verification were more likely to not achieve *Verification* and had higher *Deviance*, according to 5-out-of-6 of the correlations. Team size did not clearly matter. Finally, as should be clear by now, having more transparent replication materials mattered greatly.

What also mattered was whether my qualitative coding revealed routine researcher variability in a given team. In both the original and curated results, teams that committed routine researcher variability were significantly less likely to have *Verifications* (r = -0.081) and especially *Exact Verifications* (r = -0.343). In the raw data there were not more or less likely to have *Deviance*, but in the curated data they were as expected (r = 0.228). There is a caveat here. I had to remove 4 teams from the calculation of the curated group correlations because these teams had major mistakes but no possible counterfactual, thus they had un-curated non-routine variability in addition to routine variability and were unfit for comparison. Figure 3 decomposes the role of variability by experimental group.

Figure 3. The Role of Routine Researcher Variability in Replication Results.



Note: Deviance = absolute difference between reported odds-ratios and the original study. The percentages refer to the share of replicators from that group for which the Deviance is calculated.

Only 34% of the teams in the transparent group had routine researcher variability, but of these teams the *Deviance* was significantly higher (blue versus goldenrod bars). The opaque group had 88% of teams with researcher variability, yet it had a much lower statistical association with only a slightly larger rate of *Deviance* (green versus orange bars); but nonetheless a significantly higher rate of deviance from the original study's results suggesting it could be a cause of a lack of verifiability in this group.

4 DISCUSSION

Under the assumption that data generated from the crowdsourced replication has ecological validity as 'real-world' research, I conclude that replications are not reliable. Assuming that the two groups in this project reflect extremes in the range of transparency in replication materials found in the social science literature, the pooled average rate of 94.6% for verification misses a generous 95% cut-off, and is far from the preferred 99% cut-off. At 96.4%, it would take at least three independent replications to achieve a reliability of 95%, defined as a majority of replications in any sample of replicators verifying the original study. This is calculated with a 94.6% binomial probability of a successful verification[4]. Only three replications may sound promising, but consider that replications are quite rare and appear mostly in isolation. Simply having more single replications will not likely solve the replication crisis, even if we as scientists expected that every study had at least one replicator (Hoffmann et al. 2020; Loken and Gelman 2017). When coupled with the fact that many journals tend to avoid replications, especially ones that overturn their own previously published results, and that replications are time-consuming and often not institutionally supported these results might further reduce the appetite for replications among sociologists (Breznau 2021).

But how serious is the routine researcher variability observed in this study really? I argue that this study offers a most conservative case because a verification replication involves so few decisions to make. Thus, in more complex or decision-rich research forms the researcher degrees of freedom grow exponentially and the reliability rapidly decreases. This is evidenced by routine researcher variability causing more deviance among the opaque group (see Figure 3). The opaque group is much closer to 'just doing research' than doing a verification as they faced many small decisions to make without simply looking at the original code. Certainly analogous evidence suggests inter-researcher variation when researchers conduct similar research tasks that extend beyond verifications (Bastiaansen et al. 2020; Breznau 2016; Dutilh et al. 2019; Landy et al. 2020; Silberzahn et al. 2018). Any generalizability of these findings beyond simple analytical verifications is nearly impossible to test unfortunately, given the difficulty in obtaining a reliable prior probability of coming to certain results in any given study. Prediction markets or z-curves are suggested options to estimate plausible expected replicability rates (Dreber et al. 2015; Schimmack 2020), but any attempt to identify a 'true' replicability rate can quickly digress into a philosophical mire regarding the nature of truth.

It seems that subtleties and idiosyncrasies creep into the process of research leading to outcome variation and that this should be true in any kind of research. Under this assumption, I can tentatively generalize the impact of these findings to research outside of the verification realm where prior probabilities are lower than one. This changes the question from 'how many replicators?' to 'how many researchers are necessary to achieve reliability?'. Using the same binomial calculations suggests that

[4] The cumulative probability is calculated as P for $X > x$, where X is a variable measuring the number of verifications and x the number of successful verifications required to have a 50% verification rate in any given set of replications, in other words it is the cumulative probability of having a majority of replication results verify the original. For example, in one replication x = 1, in two replications x = 1, in three replications x = 1.5, in four replications x = 2 and so forth. Having $X > x$ means that verifications will be a majority, that they will occur in more than 50% of the replication attempts. Therefore, I need to calculate n, the number of trials to achieve at least this P value so that this majority is achieved 90% of the time. The formula for P in Bernoulli trials (where only one binomial outcome is possible) is $P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$. This iterates to three replications because for n = 1, P = 0; for n = 2, P = 0.894; for n = 3, P = 0.992.

seven researchers would be necessary to reliably come to a consensus if the prior probability were only 0.80, and seventeen if it were 0.70.[5] If this is a true picture of the state of the art in social science, it must be a factor in the ostensible 'replication crisis'. Real-world phenomena are already hidden from our observational eyes by various layers of measurement error and under-developed theory and models (Auspurg and Brüderl 2021). Researcher variability adds a meta-source of variability that is not measurement error or model misspecification in the sense that it does not derive from the act of measurement or modeling. The impact of this normally unobserved source of error in any given study is difficult to assess, because it requires really strong counterfactuals. I can only assert that overall error will be somewhat larger than researchers typically find in assessments of uncertainty in a single study or in population sampling because of routine researcher variability across researchers – a finding worthy of the methods agenda for sociologists, if not data-analytic science in general.

As a limitation, it is possible that the peculiarities of this task involving ISSP data with a 10-category employment variable and a 7-category education variable (at least in the 1996 wave) made researchers particularly prone to routine researcher variability. My only argument against this is that sociologists do a great deal of survey-based research and most surveys generate data on ISCO codes, education categories (that often vary by country) and several labor market statuses that are not always consistent (like respondents reporting being "unemployed" in one question and working "part-time" in another). Therefore, sociologists should have experienced the tasks in this replication as a type of research well within their standard paradigm.

On another note, the role of software cannot be understated. For one, it may surprise some readers that software does not always round numbers following the 0.5 cutoff. This is because numbers are stored in binary code which means that a score of exactly 0.5 could be stored as 0.499999999 and thus rounded down instead of up! Different rounding or binary defaults in software lead to variability. More striking though is the procedural aspect, what might be termed *mores* of software usage. I am left with the conclusion that Stata users are either qualitatively different, experience differences in statistical training on average or that their practice of using Stata comes with a procedure that somehow influences the results in ways that differ from other software users (see Figure 2). Otherwise, I cannot explain why, in the absence of any code or knowledge of the previous study in the opaque group, replicators using Stata were more likely to come to similar results of a study just because that study also happened to be done in Stata. This should be further explored. It leaves a sense of unease regarding potentially different research norms that associate with software, because software should just be a tool that provides specific routines

[5] At 0.80 probability, a five researcher sample achieves a 94.2% likelihood of a majority coming to a given 'correct' answer, thus it takes 7 to achieve a greater-than-95% likelihood of 96.7%; for 0.60 seven researchers only achieve a 87.3% likelihood and it would take seventeen to get above 95% using the same formula above.

for the users. In modern times software has become so sophisticated that it can ‘think’ on behalf of the users through many hidden defaults that increase with the complexity of the model, but can lead to different results as shown in one comparison in education research (McCoach et al. 2018).

5 Conclusion

The basic conclusion here is a need for transparency. The opaque group attempted to replicate under very intransparent conditions, without code and without even numerical results. Not surprisingly they were far less likely to verify the original study. This is a powerful lesson that should motivate researchers to make their workflows transparent. The risk of someone replicating a given researcher’s work and coming to false conclusions is an inverse function of transparency. Potential false conclusions could be highly damaging to the scientific process because it would cast doubt where it is not necessarily needed and lead to additional, unnecessary work. Simply put, less transparency leads to more noise in research results.

I argue that the results of this study might be applicable to conceptual replications if not research in general. Similar to qualitative research where contexts often cannot be reproduced, quantitative researchers should be prepared to admit that they also ‘cannot wade through the same river twice’ due to researcher variability. I assume that if the same participants were asked to do the same replication again but start from scratch with coding, that there would be intra-researcher variation as well. For example, some researchers might no longer have access to a paid software and switch to R, but have less experience or face alternative defaults. Others might be under more time pressure this time around and make faster choices. Others still might have adopted a new standard way of dealing with missing cases, or, just randomness that unfolds in the routine practices of doing research. Even a tightly controlled attempt at reproducing previous work still appears to have a degree of analytical flexibility.

The original study of Brady and Finnigan (2014) replicated herein found results in all directions and concluded that there was no general effect of immigration on policy preferences, thus it would take an extremely unlikely wave of positive or negative coefficients to overturn this, not something that noise from researcher variability is likely to cause. In other words, the 5.4% ‘failure’ rate in verifying their findings is of no substantial threat to their conclusions. If the crowdsourced replication were repeated, it might be more interesting to target a study with overwhelmingly positive or null test results, to see if they hold given the introduction of noise. All I can conclude from the crowdsourced replication is that the impact of researcher variability was to make any single replication effect unreliable.

In conclusion, I underscore that researchers, through very little effort, can substantively come to different results. This means that if a researcher has a motivation to find support or rejection of a hypothesis, even under controlled conditions where there should not be any research choices, it can find that support. This is one in a frenzy of warnings coming out of science in recent years about the reliability of research. We as social scientists need stronger checks in place, we cannot rely on researchers alone to perform the best research, but we must concede that some variability is beyond our control. Merton argued that sociological credibility depends on researchers arriving at similar answers (Merton 1973), but the results of this study suggest that there are limits to credibility that might simply exist in the process of research. We should be cautious in expecting consensus, depending on how serious researcher variability is in standard (non-verification) research. Sociology lags behind the behavioral sciences, economics and political science in adopting replication and transparency practices. Sociology in fact could leap ahead quickly by addressing researcher variability and transparency alongside replications which are just one small tool in the toolkit for improving the discipline.

6 REFERENCES

Auspurg, Katrin, and Josef Brüderl. 2021. "Has the Credibility of the Social Sciences Been Credibly Destroyed? Reanalyzing the 'Many Analysts, One Data Set' Project by Causal Reasoning and Multiverse Analysis."

Bastiaansen, Jojanneke A., Yoram K. Kunkels, Frank J. Blaauw, Steven M. Boker, Eva Ceulemans, Meng Chen, Sy-Miin Chow, Peter de Jonge, Ando C. Emerencia, Sacha Epskamp, Aaron J. Fisher, Ellen L. Hamaker, Peter Kuppens, Wolfgang Lutz, M. Joseph Meyer, Robert Moulder, Zita Oravecz, Harriëtte Riese, Julian Rubel, Oisín Ryan, Michelle N. Servaas, Gustav Sjobeck, Evelien Snippe, Timothy J. Trull, Wolfgang Tschacher, Date C. van der Veen, Marieke Wichers, Phillip K. Wood, William C. Woods, Aidan G. C. Wright, Casper J. Albers, and Laura F. Bringmann. 2020. "Time to Get Personal? The Impact of Researchers Choices on the Selection of Treatment Targets Using the Experience Sampling Methodology." *Journal of Psychosomatic Research* 137:110211. doi: 10.1016/j.jpsychores.2020.110211.

Brady, David, and Ryan Finnigan. 2014. "Does Immigration Undermine Public Support for Social Policy?" *American Sociological Review* 79(1):17–42. doi: 10.1177/0003122413513022.

Breznau, Nate. 2016. "Secondary Observer Effects: Idiosyncratic Errors in Small-N Secondary Data Analysis." *International Journal of Social Research Methodology* 19(3):301–18. doi: 10.1080/13645579.2014.1001221.

Breznau, Nate. 2021. "Does Sociology Need Open Science?" *Societies* 11(1):9. doi: 10.3390/soc11010009.

Breznau, Nate, Eike Mark Rinke, and Alexander Wuttke. 2018. "Pre-Registered Report for 'How Reliable Are Replications? Measuring Routine Researcher Variability in Macro-Comparative Secondary Data Analyses.'"

Christensen, Garret, Jeremy Freese, and Edward Miguel. 2019. *Transparent and Reproducible Social Science Research*. Los Angeles, Calif.: University of California Press.

Collins, Harry M. 1985. *Changing Order: Replication and Induction in Scientific Practice*. London, Beverly Hills & New Delhi: Sage Publications.

Dreber, Anna, Thomas Pfeiffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A. Nosek, and Magnus Johannesson. 2015. "Using Prediction Markets to Estimate the Reproducibility of Scientific Research." *Proceedings of the National Academy of Sciences* 112(50):15343–47. doi: 10.1073/pnas.1516179112.

Dutilh, Gilles, Jeffrey Annis, Scott D. Brown, Peter Cassey, Nathan J. Evans, Raoul P. P. P. Grasman, Guy E. Hawkins, Andrew Heathcote, William R. Holmes, Angelos-Miltiadis Krypotos, Colin N. Kupitz, Fábio P. Leite, Veronika Lerche, Yi-Shin Lin, Gordon D. Logan, Thomas J. Palmeri, Jeffrey J. Starns, Jennifer S. Trueblood, Leendert van

Maanen, Don van Ravenzwaaij, Joachim Vandekerckhove, Ingmar Visser, Andreas Voss, Corey N. White, Thomas V. Wiecki, Jörg Rieskamp, and Chris Donkin. 2019. "The Quality of Response Time Data Inference: A Blinded, Collaborative Assessment of the Validity of Cognitive Models." *Psychonomic Bulletin & Review* 26(4):1051–69. doi: 10.3758/s13423-017-1417-2.

Eubank, Nicholas. 2016. "Lessons from a Decade of Replications at the Quarterly Journal of Political Science." *PS: Political Science & Politics* 49(2):273–76. doi: DOI: 10.1017/S1049096516000196.

Freese, Jeremy, and David Peterson. 2017. "Replication in Social Science." *Annual Review of Sociology* 43(1):147–65. doi: 10.1146/annurev-soc-060116-053450.

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

Hardwicke, Tom E., Michael C. Frank, Kyle MacDonald, Erica J. Yoon, Michael Henry Tessler, Sara Altman, Bria Long, Maya B. Mathur, Gustav Nilsonne, George C. Banks, Elizabeth Clayton, Mallory C. Kidwell, Alicia Hofelich Mohr, and Richie L. Lenne. 2018. "Data Availability, Reusability, and Analytic Reproducibility: Evaluating the Impact of a Mandatory Open Data Policy at the Journal Cognition." *MetaArcXiv Preprints* Open Science Framework. doi: 10.31222/osf.io/39cfb.

Hoffmann, Sabine, Felix D. Schönbrodt, Ralf Elsas, Rory Wilson, Ulrich Strasser, and Anne-Laure Boulesteix. 2020. *The Multiplicity of Analysis Strategies Jeopardizes Replicability: Lessons Learned across Disciplines*. *preprint*. MetaArXiv.

Jacoby, William G., Sophia Lafferty-Hess, and Thu-Mai Christian. 2017. "Should Journals Be Responsible for Reproducibility? | Inside Higher Ed." *Inside Higher Ed Blog*. Retrieved July 22, 2019 (https://www.insidehighered.com/blogs/rethinking-research/should-journals-be-responsible-reproducibility).

Janz, Nicole. 2015. "Leading Journal Verifies Articles before Publication – So Far, All Replications Failed." *Political Science Replication Blog*. Retrieved July 22, 2019 (https://politicalsciencereplication.wordpress.com/2015/05/04/leading-journal-verifies-articles-before-publication-so-far-all-replications-failed/).

Landy, Justin F., Miaolei Liam Jia, Isabel L. Ding, Domenico Viganola, Warren Tierney, Anna Dreber, Magnus Johannesson, Thomas Pfeiffer, Charles R. Ebersole, and Quentin F. Gronau. 2020. "Crowdsourcing Hypothesis Tests: Making Transparent How Design Choices Shape Research Results." *Psychological Bulletin*.

Loken, Eric, and Andrew Gelman. 2017. "Measurement Error and the Replication Crisis." *Science* 355(6325):584–85. doi: 10.1126/science.aal3618.

Maxwell, Scott E., Michael Y. Lau, and George S. Howard. 2015. "Is Psychology Suffering from a Replication Crisis? What Does 'Failure to Replicate' Really Mean?" *American Psychologist* 70(6):487–98. doi: 10.1037/a0039400.

McCoach, D. Betsy, Graham G. Rifenbark, Sarah D. Newton, Xiaoran Li, Janice Kooken, Dani Yomtov, Anthony J. Gambino, and Aarti Bellara. 2018. "Does the Package Matter? A Comparison of Five Common Multilevel Modeling Software Packages:" *Journal of Educational and Behavioral Statistics*. doi: 10.3102/1076998618776348.

Merton, Robert K. 1973. *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago press.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349(6251). doi: 10.1126/science.aac4716.

Schimmack, Ulrich. 2020. "A Meta-Psychological Perspective on the Decade of Replication Failures in Social Psychology." *Canadian Psychology/Psychologie Canadienne* 61(4):364–76. doi: 10.1037/cap0000246.

Silberzahn, Raphael, Eric 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. doi: 10.1177/2515245917747646.

Stockemer, Daniel, Sebastian Koehler, and Tobias Lentz. 2018. "Data Access, Transparency, and Replication: New Insights from the Political Behavior Literature." *PS: Political Science & Politics* 51(4):799–803. doi: DOI: 10.1017/S1049096518000926.

Thompson, William Hedley, Jessey Wright, Patrick G. Bissett, and Russell A. Poldrack. 2020. "Dataset Decay and the Problem of Sequential Analyses on Open Datasets" edited by P. Rodgers, C. I. Baker, N. Holmes, C. I. Baker, and G. A. Rousselet. *ELife* 9:e53498. doi: 10.7554/eLife.53498.

Wicherts, Jelte M., Coosje L. S. Veldkamp, Hilde E. M. Augusteijn, Marjan Bakker, Robbie C. M. van Aert, and Marcel A. L. M. van Assen. 2016. "Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking." *Frontiers in Psychology* 7:1832. doi: 10.3389/fpsyg.2016.01832.

Young, Cristobal. 2018. "Model Uncertainty and the Crisis in Science." *Socius* 4:2378023117737206. doi: 10.1177/2378023117737206.

7 APPENDIX A, PARTICIPANT INSTRUCTIONS

7.1 Transparent Group Instructions

==

In this project we are crowdsourcing the replication of a 2014 study by Brady and Finnigan (B&F). The published paper and online supplemental materials are attached to this email and in a shared folder (see link below). There are many different types of replication. Your team has only one goal in this first stage of replication. That is to replicate this study to determine *verifiability*. You are to assess whether the reported results of the study follow appropriately from the data and methods employed by the original authors.

We provide you with the same two waves of *International Social Survey Program* (ISSP) data, the country-level data, and the analytical code (*Stata* format) used by B&F, and you should follow their reported methods as closely as possible to determine if:

1. their results are reproducible - to check their results;
2. the results you find (whether identical or not) confirm their reported conclusions; and
3. the methods they describe in their paper are accurately reflected in their models and results - to check their work.

We ask that you replicate their work using your preferred statistical software. That is the software that your team plans to work in throughout this entire project. This is important because there are many more stages that will build on the code you develop in this stage, and we do not expect you to learn new code for this project. We ask you to assess verifiability of their methods and results, and to do this independently of their *Stata* code (.do file), although you are welcome to use it as a guide or run it to cross-check your own code. Please do everything that the authors reported doing in executing their analyses. However, it is not necessary to run their supplemental models or analyses for now. At a minimum we ask that you replicate the results from Tables 4 and 5. If you like you can replicate other models, but we need your verifiability test for Tables 4 and 5 - otherwise the replication will be incomplete.

To ensure that you use the correct version of the ISSP data, download these datafiles from our shared data folder (they are too large to attach to an email), they are in either *Stata*, .csv, or .xls format and titled ZA2900 and ZA4700. Note that in .csv and .xls format the data contain no meta-data (i.e., no variable labels or differentiation between string and numeric) so you might need access to additional documentation. If you cannot manage to import or work with one of these formats please contact us for transferring the data into your preferred format.

[redacted] (click to access ISSP data, plus other materials left here for convenience; if you do not have HTML enabled email you may copy and paste the link at the end of this email into your browser).

Please be sure that you document all your work and that we can reproduce your results using the code you give us. Please document any cases in which you conclude that the authors' research is not verifiable in either results or the match between what they claim to do and what they actually do (i.e., points 1-3 above). Please write a short summary of your arguments supporting claims that their reported methods *do not match* their actual methods. If during this replication concerns or ideas arise for different or better analytical strategies than those employed by the original authors, this is great, but please keep them in mind for the phase after the replication when you will be asked to expand or improve upon this particular study. But for now, we ask that you do not yet run additional analyses or alternative model specifications as these might bias your task.

Results should be submitted by September 10th, 2018 to [redacted] and must include your code saved in its own language file (e.g., .do, .R, .inp, etc) and a results table in spreadsheet format. We provide an attached Excel [template link redacted] where you can fill in your results for B&F's Tables 4 and 5, but feel free to replicate their other main models if you are interested. It is not necessary to reproduce or verify their graphs for now.

We know how much time pressure you may face as a productive scholar, but we must stress the importance of completing the replication on time as the success of the project depends on starting the next phase of the CRI on time. We estimate that this exercise may take between 5 and 14 hours of working time depending very much on your own experience with the data and/or the models employed herein. Thank you for your understanding and participation in this exciting initiative. We remind you that all participants completing the CRI tasks will be co-authors on the final paper where we present the results of the study. Do not hesitate to ask if you have questions or need assistance.

7.2 Opaque Group Instructions

==

You are now asked to replicate a study to start this project. You are assigned to replicate a published study but to do so without knowing the study. We realize this may seem unusual; however, your participation is crucially important to developing deeper knowledge about replication and crowdsourcing. We kindly ask that you attempt to replicate this study to the best of your ability using only the materials we provide, and without spending time trying to 'figure out' where it came from. Again, your cooperation in this collaborative and co-authored research project is of great importance.

Attached to this email is a Methods and Results section from this study, re-written by us to render it anonymous. We ask that you focus entirely on replication and assess the verifiability of the study by:

1. replicating their exact models - to the best of your ability
2. checking if your results match the results described in the Results section

The original authors used two waves of *International Social Survey Program* (ISSP) data and a few country-level measures. We link you to these data directly in a shared data folder (they are too large to attach to an email), they are in either *Stata*, .csv, or .xls format and titled ZA2900 (ISSP 1996), ZA4700 (ISSP 2006), and L2data (for the country-level data). Note that in .csv and .xls format the data contain no meta-data (i.e., no variable labels or differentiation between string and numeric) so you might need access to additional documentation. Please work only with the data provided as it is essential to our project that all replication teams work with identical data. If you cannot manage to import or work with one of these formats please contact us for transfering the data into your preferred format.

[redacted] (click to access ISSP and country-level data, if you do not have HTML enabled email please copy and paste the link at the end of this email into your browser).

Please work in the statistical software you normally work with. We ask that you do not learn a new software in order to participate in this initiative. Please be sure that you document all your work and that we can reproduce your results using the code you give us. If you need a additional documentation (e.g., codebooks) there are two links at the end of this email, one for each ISSP wave. If during this replication concerns or ideas arise for different or better analytical strategies than employed by the original authors, please keep them in mind for the phase after the replication when you will have the chance to share them and to do them. But for now we ask that you do not yet run additional analyses or alternative model specifications as these might bias your task.

Results should be submitted by September 10th, 2018 to [redacted]. Please include your code saved in its own language file (e.g., .do, .R, .inp, etc) and a results table in spreadsheet format (.csv, .xlsx, .gsheet etc). We provide an attached Excel [template link redacted] to give you an example of the ideal 'style' of results, and if you like you can fill in your results.

We know how much time pressure you may face as a productive scholar, but we must stress the importance of completing the replication on time as the success of the project depends on starting the next phase of the CRI on time. We estimate that this exercise may take between 5 and 14 hours of working time depending very much on your own experience with the data and/or the models employed herein. Thank you for your understanding and participation in this exciting initiative. We remind you that all participants completing the CRI tasks will be co-authors on the final paper where we present the results of the study. Do not hesitate to ask if you have questions or need assistance.

7.3 Opaque Group Methods Section

Dear CRI Participant,

The following 'Methods Section' is taken from a published study this is re-written in a way that maintains identical methods, but anonymizes it from the original study. We will reveal the original study after you submit your replication results. Please note that the original paper theoretically argued and cited reasons for research choices and conducted several sensitivity analyses with country-level variables that we have purposefully omitted here. *We want you to focus on reproducing the procedure described and verifying their conclusions in your replication.* If you feel ideally that you require more information to create these models, please just use your best judgement or whatever your standard decision might with the given information. In other words, treat what is below and in the data as the 'universe' of information available to you to reconstruct this study and then do your best. Thank you again for your participation.

Your goal is to produce two tables representing the impact of *Immigrant Stock* and *Change in Immigrant Stock* on policy attitudes - reported survey responses regarding the ideal role of government in various social policies. We ask that participants use a style following our preformatted template attached to the email [redacted] for reporting results of the various models and then save in any spreadsheet format (.xls, .csv, etc) that we can easily copy and paste (for example, no .pdf files please). Please include the significance starts in addition to the z-statistics, even though both indicate the p-value, we want you to follow what is 'standard practice' in the literature and draw your conclusions from this. After producing the tables, please compare your results to the descriptive results found in the Results section below. Please indicate if you support the descriptive results in a short written summary and please share your code, including the software and version, and any other tools you incorporated in the replication of this study.

Again, all materials and data are available in the [shared data folder link redacted].

Methods

<In the following measured variables are italicized and capitalized throughout.>

Four policy attitudes are analyzed as dependent variables, taken from the *International Social Survey Program* (ISSP). These questions start with (in verbatim English), “On the whole, do you think it should or should not be the government’s responsibility to . . . ”. Then there is a module of questions from which we draw variables in the social welfare related domains of, “... provide a decent standard of living for the old” we label this *Old Age Care*, “... provide a decent standard of living for the unemployed” labeled *Unemployed*, “... reduce income differences between the rich and the poor” labeled *Reduce Income Differences*, and “... provide a job for everyone who wants one” labeled *Jobs*. Respondents chose among ordinal categories of definitely should be, probably should be, probably should not be, and definitely should not be for each. These are collapsed into a dichotomous variable where affirmative answers =1.

The main test variables are two country-level indicators of immigration as an absolute and a relative measure. The absolute measure is *Immigrant Stock* measured as percent foreign-born out of the total population, and the relative measures is *Change in Immigrant Stock* measured as the net migration number of in-migrants minus the number of out-migrants in the last year taken as a percentage of the total population. Both variables are lagged one year behind the dependent variable. Country-level variables that might otherwise influence social welfare policy attitudes are also included as *Social Welfare Expenditures* (the commonly used ‘SOCX’ variables) as a percentage of GDP and *Employment Rate* (% of active LF).

A range of individual-level variables expected to uniquely influence social welfare policy attitudes are included. These are *Female* (=1, male=0), *Age* and *Age-squared*, education categories (*Primary or less*, *Secondary* and *University or more*; with secondary as reference), and employment categories (*Part-time*, *Not active*, *Active unemployed*, and *Full-time*; with full- time as the reference category).

The ISSP data from 1996 and 2006 are pooled and all thirteen rich democratic welfare states with data for both waves are included. Models employing country and year fixed-effects to account for both the nested structure of individuals in countries and to allow for differences between time points are employed. These models are known as “two-way fixed-effects” models in the econometric literature. These models therefore have dummy variables for countries and years.

Given uncertainties in the relationships between country-level variables, different configurations are tested but all having the same individual-level variables. The main results are reported as odds-ratios and z-statistics. Models are numbered for convenience. Models 1-4 include only *Immigrant Stock*, 5-8 include *Immigrant Stock* and *Social Welfare Expenditures*, 9-12 include *Immigrant Stock* and *Employment Rate*, 13-16 include only *Change in Immigrant Stock*, 17-20 include *Change in Immigrant Stock* and *Social Welfare Expenditures*, and 21-24 include *Change in Immigrant Stock* and *Employment Rate*.

Results

In the first models (1-4) analyzing the impact of *Immigrant Stock*, odd-ratios and significance tests suggest that a one percent increase in *Immigrant Stock* statistically increases the likelihood of agreeing with *Old Age Care -* an increase significantly different from zero. It has no effect on *Unemployment*, so an impact not statistically different from zero. It statistically decreases the likelihood of agreeing with the variables *Reduce Income Differences* and *Jobs*. In the next four models including *Social Welfare Expenditures* (5-8), *Immigrant Stock* shows the exact same pattern of direction and significance across the four dependent variables. In the final four models using *Immigrant Stock* with *Employment Rate* added in (9-12) results remain the same except that *Old Age Care* drops out of significance.

Results for *Change in Immigrant Stock* alone (models 13-16) reveal that it has a statistically significant impact on increasing the likelihood of agreement with *Old Age Care* and *Jobs*, while having a not significantly different from zero impact on *Unemployment* and *Reduce Income Differences*. Models including *Social Welfare Expenditure* (17-20) do not change these results at all. However, addition of *Employment Rate* (21-24) leads to *Change in Immigrant Stock* significantly increasing the likelihood of agreement with all four dependent variables.

This study concludes that there is no systematic impact of immigration on responses to these survey questions, and this is evidence that immigration does not decrease support for the social welfare state.

8 APPENDIX B, FULL CODING RESULTS BY TEAM

Common recode variations

A	'helping family member' coded 'not in LF' (was 'part-time' in original)
B	'completed primary' coded 'secondary' (was 'primary' in original)
C	'incomplete university/tertiary' coded 'university' (was 'secondary' in original)
D	'helping family member' coded using 'hours worked per week' variable to split respondents into either 'full-time' or 'part-time'
E	'unemployed' coded as 'not in LF'
F	'student' coded as 'unemployed'
G	'housewife/-man, home maker' coded as 'unemployed'
H	Recoded 'none' or 'still in school' as missing on education
I	'helping family member' coded as 'full-time'
J	'housewife/-man, home maker' coded as 'full-time'
K	'helping family member' coded as 'missing'

- - - - - - - - - - - - - - - - - - - Transparent Group - - - - - - - - - - - - - - - - - - -

	original results		curated results			
#	exact verif. rate	deviance	exact verif. rate	deviance	Sources of Variability	Type
2	100%	0.00	100%	0.00		
3	100%	0.00	100%	0.00		
9	100%	0.00	100%	0.00		Routine
10	21%	0.08	21%	0.08	Recoded missing values to zero in each employment category; recoded missing values to zero in self-employed variable	Routine
15	100%	0.00	100%	0.00		
16	100%	0.00	100%	0.00		
17	98%	0.00	98%	0.00	It is not possible to explain seemingly random variation at the third decimal place, this team is a good example. The results are basically identical with occasional deviance up to 0.007 from original effect sizes. This must relate to rounding at different points in the routines.	Routine
19	98%	0.00	98%	0.00		
21	100%	0.00	100%	0.00		
25	8%	0.07	52%	0.04	Reported clustered SE models on accident	Non-routine, counterfactual
					Included additional indepenent variables	Non-routine, counterfactual
					Recode variation B (education categories)	Routine
29	100%	0.00	100%	0.00		
31	88%	0.00	88%	0.00	Recode variation A (employment)	Routine
					Did not recode self-employed as missing if work-status variable was missing	Routine
34	31%	0.02	31%	0.02	Did not recode nor include any individual level control variables	Non-routine, no counterfactual
36	100%	0.00	100%	0.00		
37	40%	0.01	40%	0.01	Recoded missing on income to zero, elected not to counterfactual as this is a plausible (although highly controversial) procedural step	Routine?
					Coded "Germany" as respondents in former Western Germany only	Routine
					Included N.Ireland as part of "United Kingdom"	Routine
					Recode variation C (education)	Routine
					Recode variation I & J (employment)	Routine
38	100%	0.00	100%	0.00		
39	79%	0.01	79%	0.01	Recode variation H (education)	
					Recode variation K (employment)	Routine
40	100%	0.00	100%	0.00		
41	19%	0.08	19%	0.08	Used maximum likelihood estimation	Routine
					Recoded education as 'none', 'primary' and 'secondary'	Routine
					Recode variation A (employment)	Routine
					Income variable not recoded	Non-routine, no counterfactual
42	100%	0.00	100%	0.00		
44	100%	0.00	100%	0.00		
45	88%	0.01	88%	0.01	Control variable local not defined in submitted code, appears that year dummies were left	Unknown
47	100%	0.00	100%	0.00		

- - - - - - - - - - - - - - - - - - - Transparent Group - - - - - - - - - - - - - - - - - - -

	original results		curated results			
	exact		exact			
#	verif. rate	deviance	verif. rate	deviance	Sources of Variability	Type
48	46%	0.04	46%	0.04	Recode variation A (employment)	Routine
					'Self-employed' recoded to zero if 'not in LF' or 'unemployed' scored for employment	Routine
53	15%	0.07	98%	0.00	Recoded roughly 6 thousand cases to missing via the self-employment variable recode	Routine
56	69%	0.01	69%	0.01	After several reviews, code should produce identical results, but about 5 thousand cases were dropped somewhere, probably via listwise deletion	Routine?
60	17%	0.03	42%	0.03	Included additional independent variables	Non-routine, counterfactual
61	100%	0.00	100%	0.00		
63	100%	0.00	100%	0.00		
64	100%	0.00	100%	0.00		
65	19%	0.04	71%	0.01	Forgot 2006 wave dummy	Non-routine, counterfactual
66	13%	0.04	19%	0.04	Listwise deletion by all DVs	Routine
					Did not recode nor include any individual level control variables	Non-routine, no counterfactual
					Country treated as a variance component rather than a dummy	Non-routine, no counterfactual
					One country left out of analysis	Non-routine, counterfactual
70	48%	0.04	94%	0.01	Analyzed the two waves of data (1996 & 2006) separately, curation is an average	Non-routine, no counterfactual
71	100%	0.00	100%	0.00		
72	58%	0.01	58%	0.01	Recode variation K (employment)	Routine
					Used a slightly different by country income standardization procedure	Routine
73	100%	0.00	100%	0.00		
76	96%	0.00	96%	0.00		
78	100%	0.00	100%	0.00		
82	100%	0.00	100%	0.00		

- Opaque Group -

#	original results		curated results		Sources of Variability	Type
	exact verif. rate	deviance	exact verif. rate	deviance		
1	100%	0.00	100%	0.00	Recode variation A (employment)	Routine
4	65%	0.01	65%	0.01	Listwise deletion by all DVs	Routine
					Recode variation B & C (education)	Routine
5	10%	0.09	80%	0.01	Reverse coded 1996 and 2006 as wave indicators	Non-routine, counterfactual
					Recode variation A (employment)	Routine
					Recode variation B (education)	Routine
6	53%	0.01	53%	0.01	Recode variation B (education)	Routine
					Recode variation A (employment)	Routine
7	46%	0.04	46%	0.04	Recode variation D (employment)	Routine
					Recode variation B (education)	Routine
8	55%	0.01	55%	0.01	Recode variation H (education)	Routine
					Recode variation A (employment)	Routine
11	55%	0.02	55%	0.02	Some cases dropped due to matching the (unrelated) ID variable between waves	Unclear
					Recode variation B (education)	Routine
12	10%	0.06	10%	0.06	Included N.Ireland as part of "United Kingdom"	Routine
					Recode variation E, F & G (employment)	Routine
13	80%	0.01	80%	0.01	Recode variation B & C (education)	Routine
14	50%	0.01	50%	0.01	Listwise deletion all DVs	Routine
18	80%	0.01	80%	0.01	Recode variation H (education)	Routine
					Recode variation A (employment)	Routine
20	50%	0.02	50%	0.02	Listwise deletion by all DVs	Routine
					Recode variation A (employment)	Routine
					Recode variation B (education), plus coded missing for those with 'none' on education who were a 'student' in the employment variable	Routine
22	78%	0.01	78%	0.01	Employment and education variables left in original category coding (not recoded)	Non-routine, no counterfactual
23	80%	0.01	80%	0.01	Recode variation B & C (education)	Routine
					Recode variation A (employment), and, 'less-than part time' also coded 'not in labor force	Routine
24	75%	0.01	75%	0.01	Recode variation A (employment), and, 'less-than part time' also coded 'not in labor force	Routine
26	58%	0.01	58%	0.01	Used robust estimation routine	Non-routine, no counterfactual
					Combined information from 'years of education' variable to create 'primary or less' education variable	Routine
					Recode variation A (employment)	Routine
27	13%	0.16	13%	0.16	Merging of waves done with point-and-click in SPSS, education variable recode not clear but may blur different coding schemes between the two waves	Non-routine, no counterfactual
28	83%	0.01	83%	0.01	Centered age and all country-level variables	Unclear
					Used robust clustered SEs	Non-routine, counterfactual
30	38%	0.03	38%	0.03	Recode variation A (employment), and, 'less-than part time' also coded 'not in labor force	Routine
					Listwise deletion by all DVs	Routine

- Opaque Group -

	original results		curated results			
	exact		exact			
#	verif. rate	deviance	verif. rate	deviance	Sources of Variability	Type
32	48%	0.02	48%	0.02	Recode variation B & C (education)	Routine
					Used robust clustered SEs	Non-routine, counterfactual
33	53%	0.01	53%	0.01	Recode variation H (education)	Routine
					Recode variation A (education)	Routine
35	40%	0.02	40%	0.02	Recode variation B & C (education)	Routine
					Recode variation E, F & G (employment)	Routine
43	45%	0.01	45%	0.01	Recoded 'incomplete primary' and 'primary complete' as 'secondary'	Non-routine?
					Recode variation A (employment)	Routine
46	43%	0.01	43%	0.01	Recode variation B & C (education)	Routine
					Recode variation A (employment)	Routine
					Used robust clustered SEs	Non-routine, counterfactual
49	100%	0.00	100%	0.00		
50	28%	0.02	28%	0.02	Recode variation B (education)	Routine
					Merging process resulting in only 12 countries, mislabeled and introduction of 6,000 extra cases - not fixable in a reasonable timeframe	Non-routine, no counterfactual
51	25%	0.06	13%	0.16	Using Stata for the first time, ran multilevel logit models. Did coding of data without saving, not reproducible or curatable.	Both?
52	25%	0.02	100%	0.00	Dropped Spain but included Russia	Non-routine, counterfactual
					Reported two decimal places (therefore, only two decimal places were kept after counterfactual)	Routine
					Centered age	Routine
					'Helping family member' coded as 'unemployed'	Routine
54	53%	0.01	53%	0.01	Recode variation B (education)	Routine
					Introduced roughly 6,000 cases by recoding missing to zero	Routine
55	80%	0.01	80%	0.01	Coded missing for those with 'none' on education	Routine?
57	80%	0.01	80%	0.01	Recode variation B & C (education)	Routine
58	80%	0.01	80%	0.01	Recode variation B & C (education)	Routine
59	8%	0.13	4%	0.12	Analyzed the two waves of data (1996 & 2006) separately	Unknown
62	38%	0.02	38%	0.02	Part of the data cleaning code not provided	Routine
67	0%	0.16	20%	0.05	All DVs for 2006 wave coded 0	Non-routine, counterfactual
68	43%	0.02	43%	0.02	Recode variation H (education)	Routine
					'secondary completion' recoded to 'primary' in education variable, it appears the team used 2 through 8 rather than 1 through 7 to make their recodes; same for employment variable 2 through 11	Routine?
					Rounded output to two-decimal places	Routine
69	18%	0.03	95%	0.00	Recoded two out of four of DV to zero	Non-routine, counterfactual
					Recode variation A (employment)	Routine

- Opaque Group -

	original results		curated results			
	exact		exact			
#	verif. rate	deviance	verif. rate	deviance	Sources of Variability	Type
74	15%	0.04	15%	0.04	Used multilevel models instead of two-way fixed effects,counterfactual not possible as it would require new coding with a different package or equation	Non-routine, no counterfactual
					Recode variation D (employment)	Routine
75	45%	0.02	45%	0.02	Recode variation B & C (education)	Routine
					Used maximum likelihood estimation	Routine
77	0%	0.99	63%	0.01	Reported logit coefficients instead of odds-ratios	Non-routine, counterfactual
					Recode variation H (education)	Routine
					Clustered SEs by country	Non-routine, counterfactual
79	95%	0.00	95%	0.00		
80	5%	0.95	100%	0.00	Reported logit coefficients instead of odds-ratios	Non-routine, counterfactual
81	5%	0.12	4%	0.10	Analyzed the two waves of data (1996 & 2006) separately, curation is an average	Non-routine, no counterfactual
83	73%	0.01	73%	0.01	'less than part-time' coded as 'not in labor force' for employment category	Routine
					Recode variation B (education)	Routine
84	85%	0.01	85%	0.01	Recoded education into only two, 'primary or less' and 'secondary or more'	Routine
					'helping family member', 'housewife/-man, home maker', and 'less than part-time' coded as unemployed; and 'Other/not in labor force' coded as missing	Routine
85	78%	0.01	78%	0.01	Recode variation B & C (education)	Routine
					'helping family member', 'housewife/-man, home maker', and 'less than part-time' coded as unemployed	Routine