1 Appendix A, Participant Instructions

1.1 Transparent Group (TG) 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.

1.2 Opaque Group (OG) 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.

1.3 Opaque Group (OG) 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.

2 Appendix B, Additional Tables and Figures

Table 1. Descriptive Statistics Across Three-Ecological Scenarios

		Effect-Level			Team-Level Avg.			Team-Level Dichotomy		
Variables	Measurement	TG	OG	Pooled	TG	OG	Pooled	TG	OG	Pooled
Raw Replication Results										
Cases		1,872	1,874	3,746	39	46	85	39	46	85
Directional Reproduction	same direction =1	0.957	0.893	0.925	0.957	0.887	0.918	0.795	0.652	0.718
Exact Replication	identical at two decimals =1	0.769	0.481	0.625	0.769	0.473	0.606	0.615	0.065	0.318
Replication Error	absolute difference with original	0.013	0.071	0.042	0.013	0.073	0.046	0.013	0.072	0.045
Replication Error, SD	team-level SD	0.037	0.202	0.148	0.024	0.195	0.148	0.024	0.199	0.149
Curated Replication Results										
Cases		1,872	1,875	3,747	39	46	85	39	46	85
Directional Reproduction	same direction =1	0.982	0.923	0.952	0.982	0.917	0.946	0.872	0.761	0.812
Exact Replication	identical at two decimals =1	0.839	0.566	0.702	0.839	0.556	0.683	0.641	0.130	0.365
Replication Error	absolute difference with original	0.008	0.025	0.016	0.008	0.028	0.019	0.008	0.025	0.017
Replication Error, SD	team-level SD	0.026	0.053	0.042	0.016	0.041	0.033	0.016	0.039	0.032
Trimmed Replication Results										
Cases		1,776	1,764	3,540	37	44	81	37	44	81
Directional Reproduction	same direction =1	0.963	0.920	0.941	0.963	0.914	0.936	0.838	0.682	0.753
Exact Replication	identical at two decimals =1	0.795	0.507	0.651	0.795	0.500	0.633	0.649	0.068	0.333
Replication Error	absolute difference with original	0.012	0.067	0.040	0.012	0.069	0.043	0.012	0.069	0.043
Replication Error, SD	team-level SD	0.035	0.205	0.150	0.023	0.200	0.151	0.023	0.203	0.152

NOTE: *Directional Reproduction* are same sign (including 'same' when both are not significantly different from zero, NHST, $p < 0.05$), *Exact Replication* are identical odd-ratios within 1% of original and *Replication Error* is absolute value of the difference as a ratio, between estimated odds ratio and original odds ratio. Note that team-level avg. refers to means within teams and team-level dichotomy to 95% of successful results within team = "1"; as presented in Figure 1. TG = Transparent Group with access to all materials and OG = Opaque Group with no code and less methodological information. OG Effect- and Team-Level averages are not identical because two teams had fewer models than the other teams due to mistakes leading to over-weighting when averaging the team-averages.

Table 2. Descriptive Statistics for Team-Level Variables and Qualitative Results

Independent Variables			TG	OG	Pooled
Stata		other software =0	0.67	0.63	0.65
Sociology Degree		other degrees =0	0.49	0.50	0.49
Stats-Skill		4-question scale	-0.04	0.13	0.05
SD			1.79	1.54	1.65
Difficulty		1-question (5 ordinal levels)	2.10	2.24	2.18
SD			0.79	0.77	0.77
Team Size		1-3 persons	-0.09	0.02	-0.03
SD			0.64	0.86	0.77
Qualitative Categories					
Mistake		see text	0.23	0.31	0.28
Procedural		see text	0.15	0.71	0.47
Mistake-Procedural		see text	0.21	0.21	0.21
Missing Component		see text	0.03	0.06	0.05
Interpretational		see text	0.00	0.06	0.04
Questionable Methods Competencies		see text	0.00	0.06	0.04

NOTE: T-tests comparing the two randomized groups on the variables in the shaded box reveal that there is no significant mean difference (all p-values were above 0.35). TG = Transparent Group with access to all materials and OG = Opaque Group with no code and less methodological information.

Table 6. Qualitative Categorization of Inter-Researcher Error

- *TRANSPARENT GROUP* -										
	Raw Results			Curated Results						
ID	DR	ER	RE	DR	ER	RE	Sources of Error	Category	Notes	Curate?
2	100%	100%	0.00	100%	100%	0.00				
3	100%	100%	0.00	100%	100%	0.00				
9	100%	100%	0.00	100%	100%	0.00				
10	48%	21%	0.08	100%	81%	0.01	Used the same dependent variable in a recode routine for all other dependent variables, the result being that in 2006 all different dependent variables' values were identical.	Mistake		Yes
10							In curation, the exact replication values are close but often just miss our 0.01 cut-off (it would be 100% exact replication if we allowed 0.02). We re-ran their code substituting the dataframe created by Team 3 and got 100% exact replication. Closer inspection could not reveal the source of slightly lower case numbers than the original study.	Procedural	Coded as procedural, but exact variation source unknown/ unfound by PIs	No
15	100%	100%	0.00	100%	100%	0.00				
16	100%	100%	0.00	100%	100%	0.00				
17	100%	98%	0.00	100%	98%	0.00				
19	100%	98%	0.00	100%	98%	0.00				
21	100%	100%	0.00	100%	100%	0.00				
25	83%	8%	0.07	88%	52%	0.04	Reported clustered SE models on accident	Mistake		Yes
25							Included additional indepenent variables	Mistake		Yes
25							Recode variation B (education categories)	Mistake-Procedural		No
29	100%	100%	0.00	100%	100%	0.00				
31	100%	88%	0.00	100%	88%	0.00	Recode variation A (employment)	Mistake-Procedural		No
31							Did not recode self-employed as missing if work-status variable was missing	Mistake-Procedural		No
34	98%	31%	0.02	98%	31%	0.02	Did not recode nor include any individual level control variables	Mistake		No
36	100%	100%	0.00	100%	100%	0.00				

NOTE: Recode variations listed in Table 4, see main text. DR = "Direct Replication", ER = "Exact Replication", RE = "Replication Error". Curate? Measures whether there was a clear counterfactual and the principal investigators could infer what the authors would have done otherwise.

Table 6 Continued. Qualitative Categorization of Inter-Researcher Error

- ***TRANSPARENT GROUP*** -

	Raw Results			Curated Results						
ID	DR	ER	RE	DR	ER	RE	Sources of Error	Category	Notes	Curate?
37	100%	40%	0.01	100%	40%	0.01	Recoded missing on income to zero, elected not to counterfactual as this is a plausible (although highly controversial) procedural step	Procedural		No
37							Coded "Germany" as respondents in former Western Germany only	Mistake-Procedural		No
37							Included N.Ireland as part of "United Kingdom"	Procedural		No
37							Recode variation C (education)	Mistake-Procedural		No
37							Recode variation I & J (employment)	Procedural		No
38	100%	100%	0.00	100%	100%	0.00				
39	100%	79%	0.01	100%	79%	0.01	Recode variation H (education)	Mistake-Procedural		No
39							Recode variation K (employment)	Mistake-Procedural		No
40	100%	100%	0.00	100%	100%	0.00				
41	90%	19%	0.08	90%	19%	0.08	Used maximum likelihood estimation	Procedural	Missing parts of workflow	No
41							Recoded education as 'none', 'primary' and 'secondary'	Mistake		No
41							Recode variation A (employment)	Mistake-Procedural		No
41							Income variable not recoded, cannot guess how they would have recoded it	Mistake		No
41							Used a different country as the reference dummy category	Mistake	They specifically stated in their notes that the original study did not mention which dummy variable it used, but it was in the code	No
42	100%	100%	0.00	100%	100%	0.00				
44	100%	100%	0.00	100%	100%	0.00				
45	98%	88%	0.01	98%	88%	0.01	Control variable local not defined in submitted code, thus year dummies were left out of analysis	Mistake	Unclear if or how they would have set up their local differently	No

Table 6 Continued. Qualitative Categorization of Inter-Researcher Error

- ***TRANSPARENT GROUP*** -

	Raw Results			Curated Results						
ID	DR	ER	RE	DR	ER	RE	Sources of Error	Category	Notes	Curate?
47	100%	100%	0.00	100%	100%	0.00				
48	81%	46%	0.04	81%	46%	0.04	Recode variation A (employment)	Mistake-Procedural		No
48							'Self-employed' recoded to zero if 'not in LF' or 'unemployed' scored for employment	Mistake-Procedural		No
53	81%	15%	0.07	100%	98%	0.00	Recoded roughly 6 thousand cases to missing via the self-employment variable recode	Mistake		Yes
53										
56	100%	69%	0.01	100%	69%	0.01	Coded cases as missing rather than 0 when constructing dichotomous variables resulting in roughly 4k less cases in the analyzed data	Mistake-Procedural		No
60	92%	17%	0.03	90%	42%	0.03	Included additional independent variables	Mistake		Yes
61	100%	100%	0.00	100%	100%	0.00				
63	100%	100%	0.00	100%	100%	0.00				
64	100%	100%	0.00	100%	100%	0.00				
65	73%	19%	0.04	100%	71%	0.01	Forgot 2006 wave dummy	Mistake		Yes
66	88%	13%	0.04	85%	19%	0.04	Listwise deletion by all DVs	Procedural	This is not clearly a mistake as it could be seen as a best practice or procedural step	No
66							Did not recode nor include any individual level control variables	Mistake	Unclear how the team would have done the recodes otherwise	No
66							Included a respondent ID variable as a random-intercept	Mistake		Yes
66							One country left out of analysis	Mistake		Yes
70	100%	100%	0.00	100%	100%	0.00				
71	100%	100%	0.00	100%	100%	0.00				

Table 6 Continued. Qualitative Categorization of Inter-Researcher Error

- ***TRANSPARENT GROUP*** -

	Raw Results			Curated Results						
ID	DR	ER	RE	DR	ER	RE	Sources of Error	Category	Notes	Curate?
72	100%	58%	0.01	100%	58%	0.01	Recode variation K (employment)	Mistake-Procedural		No
72							Used a slightly different by-country income standardization procedure	Procedural		No
73	100%	100%	0.00	100%	100%	0.00				
76	100%	96%	0.00	100%	96%	0.00				
78	100%	100%	0.00	100%	100%	0.00				
82	100%	100%	0.00	100%	100%	0.00				

Table 6 Continued. Qualitative Categorization of Inter-Researcher Error

- ***OPAQUE GROUP*** -

	Raw Results			Curated Results						
ID	DR	ER	RE	DR	ER	RE	Sources of Error	Category	Notes	Curate?
1	100%	100%	0.00	100%	100%	0.00	Recode variation A (employment)	Procedural		No
4	100%	65%	0.01	100%	65%	0.01	Listwise deletion by all DVs	Procedural		No
4							Recode variation B & C (education)	Procedural		No
5	33%	10%	0.09	100%	80%	0.01	Reverse coded 1996 and 2006 as wave indicators	Mistake		Yes
5							Recode variation A (employment)	Procedural		No
5							Recode variation B (education)	Procedural		No
6	100%	53%	0.01	100%	53%	0.01	Recode variation B (education)	Procedural		No
6							Recode variation A (employment)	Procedural		No
7	92%	46%	0.04	92%	46%	0.04	Recode variation D (employment)	Procedural		No
7							Recode variation B (education)	Procedural		No
8	100%	55%	0.01	100%	55%	0.01	Recode variation H (education)	Procedural		No
8							Recode variation A (employment)	Procedural		No
11	90%	55%	0.02	90%	55%	0.02	Recode variation B (education)	Procedural		No

Table 6 Continued. Qualitative Categorization of Inter-Researcher Error

OPAQUE GROUP

	Raw Results			Curated Results						
ID	DR	ER	RE	DR	ER	RE	Sources of Error	Category	Notes	Curate?
12	80%	10%	0.06	80%	10%	0.06	Included N.Ireland as part of "United Kingdom"	Procedural		No
12							Used a different 13 country-sample based on data availability at the country-level and highest 13 scores on the social spending indicator, rather than select based on the individual-level data. In the end this led to an unbalanced time-series with some of the 13 countries having only one case in either 1996 or 2006.	Interpretational	Instructions were not 100% clear on whether to exclude country cases based on individual or country-level data	No
12							Recode variation E, F & G (employment)	Procedural		No
13	100%	80%	0.01	100%	80%	0.01	Recode variation B & C (education)	Procedural		No
14	100%	50%	0.01	100%	50%	0.01	Listwise deletion all DVs	Procedural		No
18	100%	80%	0.01	100%	80%	0.01	Recode variation H (education)	Procedural		No
18							Recode variation A (employment)	Procedural		No
20	88%	50%	0.02	88%	50%	0.02	Listwise deletion by all DVs	Procedural		No
20							Recode variation A (employment)	Procedural		No
20							Recode variation B (education), plus coded missing for those with 'none' on education who were a 'student' in the employment variable	Procedural		No
22	100%	78%	0.01	100%	78%	0.01	Employment and education variables left in original category coding (not recoded)	Mistake	Unclear how they would have recoded otherwise	No
23	100%	80%	0.01	100%	80%	0.01	Recode variation B & C (education)	Procedural		No
23							Recode variation A (employment), and, 'less-than part time' also coded 'not in labor force	Procedural		No
24	100%	75%	0.01	100%	75%	0.01	Recode variation A (employment), and, 'less-than part time' also coded 'not in labor force	Procedural		No
26	100%	58%	0.01	100%	58%	0.01	Used robust estimation routine	Mistake-Procedural	Robust estimation not mentioned in methods description, but could be a reasonable assumption given the type of model	No
26							Combined information from 'years of education' variable to create 'primary or less' education variable	Procedural		No
26							Recode variation A (employment)	Procedural		No

Table 6 Continued. Qualitative Categorization of Inter-Researcher Error

- ***OPAQUE GROUP*** -

	Raw Results			Curated Results						
ID	DR	ER	RE	DR	ER	RE	Sources of Error	Category	Notes	Curate?
27	43%	13%	0.16	43%	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	Mistake-Procedural	Missing parts of workflow	No
28	95%	83%	0.01	95%	83%	0.01	Centered age and all country-level variables	Interpretational	Not mentioned in Methods Section but plausible choice given the data and goals	No
28							Used robust clustered SEs	Mistake-Procedural	Robust estimation not mentioned in Methods Section, but could be a reasonable assumption given the type of model	No
30	100%	38%	0.03	100%	38%	0.03	Recode variation A (employment), and, 'less-than part time' also coded 'not in labor force	Procedural		No
30							Listwise deletion by all DVs	Procedural		No
30							Software accuracy	Procedural	Some variables' values had to be truncated because they were otherwise too precise for MLWin	No
32	100%	48%	0.02	100%	48%	0.02	Recode variation B & C (education)	Procedural		No
32							Used robust clustered SEs	Mistake-Procedural	Robust estimation not mentioned in Methods Section, but could be a reasonable assumption given the type of model	No
33	100%	53%	0.01	100%	53%	0.01	Recode variation H (education)	Procedural		No
33							Recode variation A (education)	Procedural		No
35	93%	40%	0.02	93%	40%	0.02	Recode variation B & C (education)	Procedural		No
35							Recode variation E, F & G (employment)	Procedural		No

Table 6 Continued. Qualitative Categorization of Inter-Researcher Error

OPAQUE GROUP

	Raw Results			Curated Results						
ID	DR	ER	RE	DR	ER	RE	Sources of Error	Category	Notes	Curate?
43	100%	45%	0.01	100%	45%	0.01	Recoded 'incomplete primary' and 'primary complete' as 'secondary'	Mistake	We hesitated to correct this given that so many other teams made so many different choices in this recode	No
43							Recode variation A (employment)	Procedural		No
46	100%	43%	0.01	100%	43%	0.01	Recode variation B & C (education)	Procedural		No
46							Recode variation A (employment)	Procedural		No
46							Used robust clustered SEs	Mistake-Procedural	Robust estimation not mentioned in Methods Section, but could be a reasonable assumption given the type of model	No
49	100%	100%	0.00	100%	100%	0.00				
50	100%	28%	0.02	100%	28%	0.02	Recode variation B (education)	Procedural		No
50							Merging process resulting in only 12 countries, mislabeled cases and thus introduced 6,000 extra cases.	Mistake	Curation would require writing nearly new code, requiring us to make too many assumptions about what they would have done	No
51	63%	13%	0.16	63%	13%	0.16	Using Stata for the first time, ran multilevel logit models. Did coding of data without saving, not reproducible or curatable.	Mistake	Missing parts of workflow; questionable methodological competencies	No
52	90%	25%	0.02	100%	100%	0.00	Dropped Spain but included Russia	Mistake		Yes
52							Reported two decimal places (therefore, only two decimal places were kept after counterfactual)	Procedural		No
52							Centered age	Interpretational	Not mentioned in Methods Section but plausible choice given data and goals	No
52							'Helping family member' coded as 'unemployed'	Procedural		No

Table 6 Continued. Qualitative Categorization of Inter-Researcher Error

- ***OPAQUE GROUP*** -

	Raw Results			Curated Results						
ID	DR	ER	RE	DR	ER	RE	Sources of Error	Category	Notes	Curate?
54	100%	53%	0.01	100%	53%	0.01	Recode variation B (education)	Procedural		No
54							Introduced roughly 6,000 cases by recoding missing to zero	Mistake	Unclear from code how they would have done it differently	No
55	100%	80%	0.01	100%	80%	0.01	Coded missing for those with 'none' on education	Mistake-Procedural	Plausible to drop these cases	No
57	100%	80%	0.01	100%	80%	0.01	Recode variation B & C (education)	Procedural		No
57							Listwise deletion on all analytical variables	Procedural		
58	100%	80%	0.01	100%	80%	0.01	Recode variation B & C (education)	Procedural		No
59	40%	8%	0.13	40%	8%	0.13	Analyzed the two waves of data (1996 & 2006) separately, results are an average	Mistake	Questionable methodoological competencies	No
62	98%	38%	0.02	98%	38%	0.02	Parts of code missing	Procedural	Missing parts of workflow	No
67	60%	0%	0.16	58%	20%	0.05	All DVs for 2006 wave coded 0	Mistake		Yes
68	98%	43%	0.02	98%	43%	0.02	Recode variation H (education)	Procedural		No
68							'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	Mistake-Procedural	Not clear how they would have done it differently	No
68							Rounded output to two-decimal places	Procedural		No
69	85%	18%	0.03	100%	95%	0.00	Recoded DV missing values to zero	Mistake		Yes
69							Recode variation A (employment)	Procedural		No
74	85%	15%	0.04	85%	15%	0.04	Used multilevel models instead of 'two-way fixed effects'	Mistake	Counterfactual not possible as it would require new coding with a different package or equation	No
74							Recode variation D (employment)	Procedural		No
75	98%	45%	0.02	98%	45%	0.02	Recode variation B & C (education)	Procedural		No
75							Used maximum likelihood estimation	Procedural	The estimator was not clearly defined, but implied from the description	No

Table 6 Continued. Qualitative Categorization of Inter-Researcher Error

OPAQUE GROUP

	Raw Results			Curated Results						
ID	DR	ER	RE	DR	ER	RE	Sources of Error	Category	Notes	Curate?
77	73%	0%	0.99	100%	63%	0.01	Reported logit coefficients instead of odds ratios	Mistake		Yes
77							Recode variation H (education)	Procedural		No
77							Clustered SEs by country	Mistake-Procedural	Robust estimation not mentioned in methods description, but could be a reasonable assumption given the type of model	No
79	100%	95%	0.00	100%	95%	0.00				
80	73%	5%	0.95	100%	100%	0.00	Reported logit coefficients instead of odds ratios	Mistake		Yes
81	53%	5%	0.12	53%	5%	0.12	Analyzed the two waves of data (1996 & 2006) separately, results are taken as the average	Mistake	Questionable methodological competencies	No
83	100%	73%	0.01	100%	73%	0.01	'less than part-time' coded as 'not in labor force' for employment category	Mistake-Procedural		No
83							Recode variation C (education)	Procedural		No
84	100%	85%	0.01	100%	85%	0.01	Recoded education into only two, 'primary or less' and 'secondary or more'	Mistake	The description was clear on this, but we do not have a counterfactual	No
84							'helping family member', 'housewife/-man, home maker', and 'less than part-time' coded as unemployed; and 'Other/not in labor force' coded as missing	Mistake-Procedural	Could arge that these are plausible steps given the ambiguities in the methods description	No
85	100%	78%	0.01	100%	78%	0.01	Recode variation B & C (education)	Procedural		No
85							'helping family member', 'housewife/-man, home maker', and 'less than part-time' coded as unemployed	Procedural		No

Table 8. Multivariate Analysis of Computational Reproducibility in 85 Teams, Curated Results

	Directional Reproduction			Exact Replication			Replication Error		
Variable	Pooled	TG	OG	Pooled	TG	OG	Pooled	TG	OG
(Intercept)	0.56^{***}	0.76^{***}	0.58^{***}	0.93^{***}	0.98^{***}	0.93^{***}	0.03^{***}	0.01^{**}	0.02^{*}
	(0.05)	(0.07)	(0.07)	(0.02)	(0.01)	(0.03)	(0.01)	(0.00)	(0.01)
Stata	0.08	0.11	0.04	0.01	0.00	0.01	0.00	-0.01	0.01
	(0.07)	(0.08)	(0.08)	(0.03)	(0.02)	(0.04)	(0.01)	(0.01)	(0.01)
Stat-skill	-0.02	-0.03	-0.01	-0.01	-0.01	-0.01	0.00	0.00	0.00
	(0.02)	(0.02)	(0.03)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)
Difficult	-0.13^{**}	-0.09	-0.16^{**}	-0.06^{***}	-0.00	-0.09^{***}	0.02^{***}	0.00	0.04^{***}
	(0.04)	(0.07)	(0.05)	(0.02)	(0.01)	(0.02)	(000)	(0.00)	(0.01)
Sociology degree	-0.00			-0.00			-0.00		
	(0.06)			(0.03)			(0.01)		
TG	0.22^{***}			0.04			-0.02^{*}		
	(0.05)			(0.02)			(0.01)		
Observations	85	39	46	85	39	46	85	39	46
R^2	0.336	0.126	0.248	0.232	0.048	0.293	0.313	0.096	0.378
R^2 adjusted	0.283	0.051	0.194	0.183	0.000	0.242	0.270	0.018	0.333

* $p<0.05$ ** $p<0.01$ *** $p<0.001$

NOTE: Directional Reproduction and Exact Replication are linear probability models and Replication Error are OLS regressions. Unstandardized OLS regression coefficients predicting outcomes aggregated to their mean by team; standard errors in parentheses. Degree omitted from group-specific regressions due to low predictive power and smaller sample sizes. TG = Transparent Group with access to all materials and OG = Opaque Group with no code and less methodological information.

Table 9. Multivariate Analysis of Computational Reproducibility in 81 Teams, Trimmed Results

	Directional Reproduction			Exact Replication			Replication Error		
Variable	Pooled	TG	OG	Pooled	TG	OG	Pooled	TG	OG
(Intercept)	0.48***	0.63***	0.53***	0.92***	0.91***	0.95***	0.45	0.03***	0.56
	(0.07)	(0.08)	(0.08)	(0.03)	(0.03)	(0.04)	(0.24)	(0.01)	(0.38)
Stata	0.13	0.22*	-0.01	0.03	0.08*	-0.04	-0.09	-0.02*	-0.33
	(0.08)	(0.10)	(0.10)	(0.03)	(0.03)	(0.05)	(0.29)	(0.01)	(0.48)
Stat-skill	-0.03	-0.06*	-0.00	-0.01	-0.01	-0.01	-0.07	0.00	-0.14
	(0.02)	(0.03)	(0.03)	(0.01)	(0.01)	(0.02)	(0.07)	(0.00)	(0.15)
Difficult	-0.13**	-0.16*	-0.12*	-0.05*	-0.04	-0.05	-0.20	0.01*	-0.34
	(0.05)	(0.08)	(0.06)	(0.02)	(0.03)	(0.03)	(0.17)	(0.01)	(0.29)
Sociology degree	-0.07			-0.03			0.03		
	(0.07)			(0.03)			(0.27)		
TG	0.25***			0.04			-0.34		
	(0.07)			(0.03)			(0.24)		
Observations	81	37	44	81	37	44	81	37	44
R^2	0.290	0.267	0.110	0.105	0.242	0.069	0.050	0.239	0.049
R^2 adjusted	0.243	0.200	0.043	0.045	0.173	0.000	0.000	0.229	0.000

* $p<0.05$ ** $p<0.01$ *** $p<0.001$

NOTE: Two teams from each group with the lowest rate of directional reproduction were trimmed. Directional Reproduction and Exact Replication are linear probability models and Replication Error are OLS regressions. Unstandardized OLS regression coefficients predicting outcomes aggregated to their mean by team; standard errors in parentheses. Degree omitted from group-specific regressions due to low predictive power and smaller sample sizes. TG = Transparent Group with access to all materials and OG = Opaque Group with no code and less methodological information

Table 11. Correlations by Group. Team-Level N = 85

	Variable	Directional Reproduction,	Exact Replication, Raw	Replication Error, Raw	Directional Reproduction,	Exact Replication,	Replication Error, Curated	Directional Reproduction,	Exact Replication,	Replication Error, Trimmed	Stata	Stat-Skill	Difficult	Sociology	Number in Team	Mistake	Procedural	Mistake-Procedural	Interpret-ational	Questionable Methods
Transparent Group	Dir Rep	1.00																		
	Exact Rep	0.75	1.00																	
	Rep Error	-0.86	-0.88	1.00																
	Dir Rep, C	NA	NA	NA	1.00															
	Exact Rep, C	NA	NA	NA	0.71	1.00														
	Rep Error, C	NA	NA	NA	-0.86	-0.89	1.00													
	Dir Rep, T	NA	NA	NA	NA	NA	NA	1.00												
	Exact Rep, T	NA	NA	NA	NA	NA	NA	0.70	1.00											
	Rep Error, T	NA	NA	NA	NA	NA	NA	-0.85	-0.89	1.00										
	Stata	0.04	0.24	-0.18	0.36	0.37	-0.38	0.38	0.35	-0.39	1.00									
	Stat-Skill	-0.21	-0.13	0.18	-0.17	-0.19	0.19	-0.16	-0.21	0.19	0.06	1.00								
	Difficult	0.01	-0.20	0.13	-0.22	-0.27	0.27	-0.24	-0.27	0.28	-0.17	-0.31	1.00							
	Sociology	0.13	0.14	-0.15	0.19	0.08	-0.09	0.14	0.01	-0.04	0.36	-0.03	-0.14	1.00						
	# in Team	-0.16	-0.07	0.07	-0.24	-0.15	0.20	-0.19	-0.07	0.14	0.02	-0.11	-0.30	0.14	1.00					
	Mistake	-0.65	-0.72	0.77	-0.50	-0.78	0.74	-0.55	-0.79	0.83	-0.38	0.13	0.29	-0.19	0.07	1.00				
	Procedural	-0.13	-0.54	0.33	-0.21	-0.43	0.29	-0.20	-0.40	0.28	-0.17	-0.12	0.28	-0.04	0.02	0.27	1.00			
	Mistake-Pr.	-0.33	-0.48	0.45	-0.05	-0.31	0.22	0.05	-0.31	0.16	0.17	0.29	0.03	0.17	0.09	0.04	0.42	1.00		
	Interpret.	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
	Questionable	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Table 11. Con’t. Correlations by Group. Team-Level N = 85

	Variable	Directional Reproduction,	Exact Replication, Raw	Replication Error, Raw	Directional Reproduction,	Exact Replication,	Replication Error, Curated	Directional Reproduction,	Exact Replication,	Replication Error, Trimmed	Stata	Stat-Skill	Difficult	Sociology	Number in Team	Mistake	Procedural	Mistake-Procedural	Interpret-ational	Questionable Methods
Opaque Group	Dir Rep	1.00																		
	Exact Rep	0.73	1.00																	
	Rep Error	-0.90	-0.71	1.00																
	Dir Rep, C	NA	NA	NA	1.00															
	Exact Rep, C	NA	NA	NA	0.75	1.00														
	Rep Error, C	NA	NA	NA	-0.25	-0.38	1.00													
	Dir Rep, T	NA	NA	NA	NA	NA	NA	1.00												
	Exact Rep, T	NA	NA	NA	NA	NA	NA	0.74	1.00											
	Rep Error, T	NA	NA	NA	NA	NA	NA	-0.32	-0.40	1.00										
	Stata	0.11	0.17	-0.06	0.11	0.17	-0.06	-0.06	0.09	-0.07	1.00									
	Stat-Skill	0.07	0.11	-0.11	0.07	0.11	-0.11	-0.01	0.07	-0.12	0.16	1.00								
	Difficult	-0.37	-0.40	-0.12	-0.37	-0.40	-0.12	-0.22	-0.33	-0.12	-0.39	-0.29	1.00							
	Sociology	0.04	0.13	0.00	0.04	0.13	0.00	-0.10	0.07	-0.01	0.59	0.37	-0.41	1.00						
	# in Team	0.10	0.28	0.07	0.10	0.28	0.07	0.18	0.32	0.07	0.18	0.04	-0.41	0.26	1.00					
	Mistake	-0.49	-0.56	0.32	-0.49	-0.56	0.32	-0.56	-0.57	0.33	-0.14	-0.25	0.41	-0.14	-0.28	1.00				
	Procedural	0.29	0.05	-0.18	0.29	0.05	-0.18	0.17	-0.05	-0.20	0.27	0.26	-0.35	0.12	0.06	-0.42	1.00			
	Mistake-Pr.	0.02	0.08	0.15	0.02	0.08	0.15	0.14	0.13	0.16	-0.03	0.01	0.02	0.00	0.18	-0.14	-0.18	1.00		
	Interpret.	-0.03	-0.09	-0.06	-0.03	-0.09	-0.06	-0.07	-0.12	-0.06	0.02	0.32	-0.07	0.26	0.03	0.00	0.04	0.07	1.00	
	Questionable	-0.50	-0.39	-0.03	-0.50	-0.39	-0.03	-0.45	-0.34	-0.03	0.02	-0.19	0.45	-0.09	-0.20	0.38	-0.36	-0.14	-0.07	1.00