Supplementary Materials for
Investigating Variation in Replicability: A "Many Labs" Replication Project

Richard A. Klein[1], Kate A. Ratliff[1], Michelangelo Vianello[2], Reginald B. Adams, Jr.[3], Štěpán Bahník[4], Michael J. Bernstein[5], Konrad Bocian[6], Mark J. Brandt[7], Beach Brooks[1], Claudia Chloe Brumbaugh[8], Zeynep Cemalcilar[9], Jesse Chandler[10, 37], Winnee Cheong[11], William E. Davis[12], Thierry Devos[13], Matthew Eisner[14], Natalia Frankowska[6], David Furrow[15], Elisa Maria Galliani[2], Fred Hasselman[16, 38], Joshua A. Hicks[12], James F. Hovermale[17], S. Jane Hunt[18], Jeffrey R. Huntsinger[19], Hans IJzerman[7], Melissa-Sue John[20], Jennifer A. Joy-Gaba[17], Heather Barry Kappes[21], Lacy E. Krueger[18], Jaime Kurtz[22], Carmel A. Levitan[23], Robyn Mallett[19], Wendy L. Morris[24], Anthony J. Nelson[3], Jason A. Nier[25], Grant Packard[26], Ronaldo Pilati[27], Abraham M. Rutchick[28], Kathleen Schmidt[29], Jeanine L. Skorinko[20], Robert Smith[14], Troy G. Steiner[3], Justin Storbeck[8], Lyn M. Van Swol[30], Donna Thompson[15], A. E. van 't Veer[7, 31], Leigh Ann Vaughn[32], Marek Vranka[33], Aaron L. Wichman[34], Julie A. Woodzicka[35], Brian A. Nosek[29, 36]

[1]Department of Psychology, University of Florida, Gainesville, FL 32611; [2]Department FISPPA, Applied Psychology, University of Padua, 35131 Padua, Italy; [3]Department of Psychology, The Pennsylvania State University, University Park, PA 16802; [4]Department of Psychology II, Social Psychology, University of Würzburg, Würzburg, Germany; [5]Department of Psychology, Pennsylvania State University Abington, Abington, PA 19001; [6]Department of Psychology, University of Social Sciences and Humanities Campus Sopot, Sopot, Poland; [7]Department of Social Psychology, Tilburg University, P.O. Box 90153, Tilburg, 5000 LE, Netherlands; [8]Department of Psychology, Queens College, City University of New York, New York, NY 11367; [9]Department of Psychology, Koç University, 34450 Istanbul, Turkey; [10]Institute for Social Research, University of Michigan, Ann Arbor, MI 48109; [11]Department of Psychology, HELP University, 50490 Kuala Lumpur, Malaysia; [12]Department of Psychology, Texas A&M University, College Station, TX 77843; [13]Department of Psychology, San Diego State University, San Diego, CA 92182; [14]Fisher College of Business, Ohio State University, Columbus, OH 43210; [15]Department of Psychology, Mount Saint Vincent University, Nova Scotia, Canada; [16]Behavioral Science Institute, Radboud University Nijmegen, Nijmegen, Netherlands; [17]Department of Psychology, Virginia Commonwealth University, Richmond, VA 23284; [18]Department of Psychology, Counseling, and Special Education, Texas A&M University-Commerce, Commerce, TX 75429; [19]Department of Psychology, Loyola University Chicago, Chicago, IL 60626; [20]Department of Psychology, Worcester Polytechnic Institute, Worcester, MA 01609; [21]Department of Management, London School of Economics and Political Science, London WC2A 2AE, UK; [22]Department of Psychology, James Madison University, Harrisonburg, VA 22807; [23]Department of Cognitive Science, Occidental College, Los Angeles, CA 90041; [24]Department of Psychology, McDaniel College, Westminster, MD 21157; [25]Psychology Department, Connecticut College, New London, CT 06320; [26]School of Business & Economics, Wilfrid Laurier University, Waterloo, ON, Canada; [27]Social and Work Psychology Department, University of Brasilia, DF, Brazil; [28]Department of Psychology, California State University, Northridge, Northridge, CA 91330; [29]Department of Psychology, University of Virginia, Charlottesville, VA 22904; [30]Department of Communication Arts, University of Wisconsin-Madison, Madison, WI 53706; [31]TIBER (Tilburg Institute for Behavioral Economics Research), Tilburg University, P.O. Box 90153, Tilburg, 5000 LE, Netherlands; [32]Department of Psychology, Ithaca College, Ithaca, NY 14850; [33]Department of Psychology, Charles University, Prague, Czech Republic; [34]Psychological Sciences Department, Western Kentucky University, Bowling Green, KY 42101; [35]Department of Psychology, Washington and Lee University, Lexington, VA 24450; [36]Center for Open Science, Charlottesville, VA 22903; [37]PRIME Research, Ann Arbor, MI; [38]School of Pedagogical and Educational Science, Radboud University Nijmegen, Nijmegen, Netherlands

Method

After the 12 studies, all participants completed an instructional manipulation check (Oppenheimer et al., 2009). In the instructional manipulation check, participants are presented with a question that contains an instruction to ignore the question asked and instead to click the header at the top of the screen. Participants who click the title at the top of the screen are considered to pass the Instructional Manipulation Check (IMC). Oppenheimer et al. (2009) found that 54% of participants passed their version of the IMC. We chose to include the IMC because it may offer a quick way for experimenters to increase the power of their experiments. Including it in this replication allows us to test whether it provides a benefit across a variety of studies and contexts, particularly whether contexts with high IMC failure rates show less replicability of findings than those with low failure rates. We are administering the IMC after the 12 studies to ensure we do not unintentionally manipulate participant attention to detail. Participants then completed a demographics questionnaire and responded to some supplemental questions relating to the flag-priming study. They were debriefed at the conclusion of the study.

Data analyses and differences from the pre-registered proposal.

6344 valid and complete study sessions were collected across 36 samples. Confirmatory tests have been conducted exactly as described in the pre-proposal, unless specifically noted.

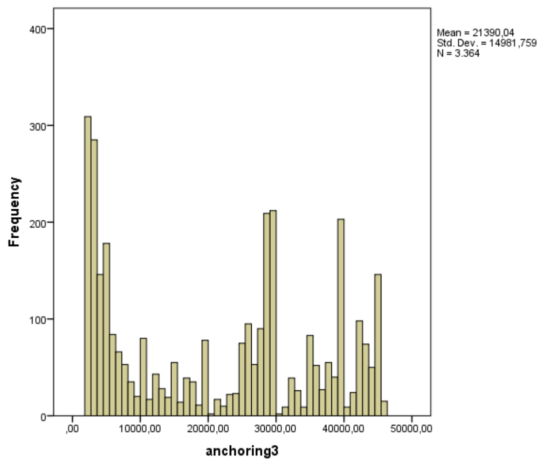
1 *Sunk costs (Oppenheimer et al., 2009).* In the Oppenheimer et al. (2009) demonstration of sunk costs, participants were slightly more likely to go to the game if they had paid for the ticket (M = 7.46, SD not reported) than if the ticket had been free (M = 6.93, SD not reported), F(1, 211) = 2.74, p = .1, partial η^2=.01. To test for replication we conducted an independent samples t-test with condition (paid ticket, free ticket) as the independent variable and likelihood of going to the game as the dependent variable. In the replication study participants were more likely to go to the game if they had paid for the ticket (M = 7.85, SD = 2.02) than if the ticket had been free (M = 7.24, SD = 2.42), t(6328) = 10.83, p <.001, d=.27. 14 subjects failed to provide a valid answer.

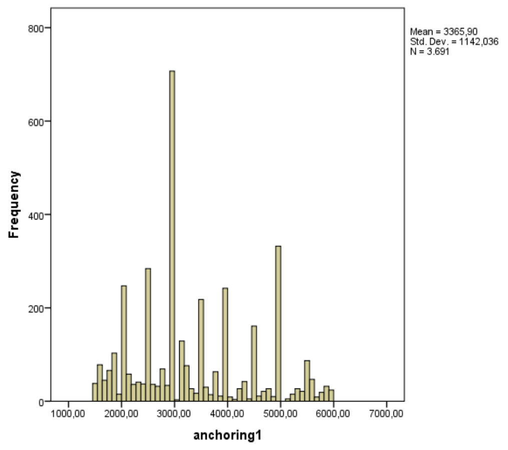
2 *Gain versus loss framing for combating disease (Tversky & Kahneman, 1981).* In the original study, 72% of participants in Condition 1 chose to adopt Program A (saving 200 people) and 28% chose Program B (a 1/3 probability that 600 people will be saved). In Condition 2 this effect was reversed, as Program C (400 people die) was adopted by only 22 percent while 78 percent selected Program D (1/3 probability that no one will die). In the replicated study, 65.5% of participants in Condition 1 chose to adopt Program A (saving 200 people) and 36.8% chose Program B (a 1/3 probability that 600 people will be saved). In Condition 2 this effect was reversed, as Program C (400 people die) was adopted by only 34.5% percent while 63.2% percent selected Program D (1/3 probability that no one will die). To test for replication we conducted a chi-square analysis on program choice with condition (gain-frame or loss-frame) as a between-subjects variable: $\chi^2(1) = 516.41$, p < .0001, d = .60. 73 subjects failed to provide a valid answer.

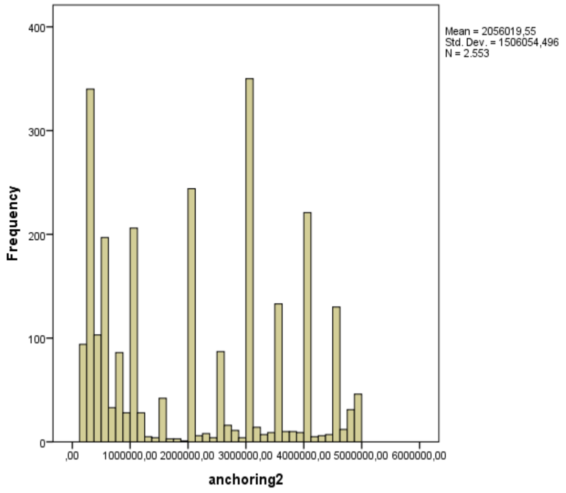
3 *Anchoring (Jacowitz & Kahneman, 1995).* In the original study, a point-biserial correlation was computed between subjects' estimates and the anchor they had seen. The mean point-biserial correlation

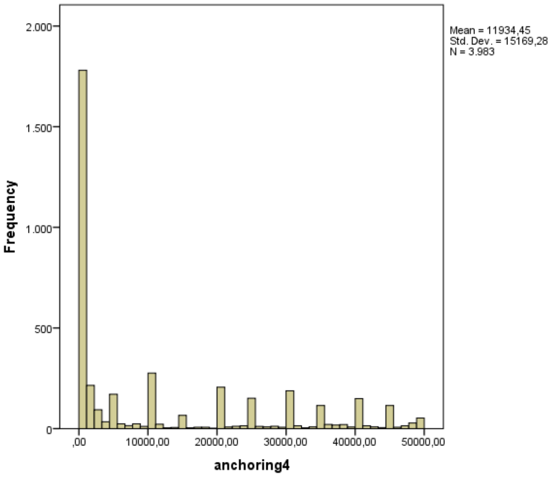
over the 15 topics was .42. This study included two exclusion rules for open response data: (1) responses must be interpretable as a number, and (2) all responses outside of the anchor end points will be converted to missing data (e.g., responses greater than 5 million or less than 200,000 for the Chicago population item). Hence directly interpretable text responses have been transformed into numbers (e.g., the string " million" have been substituted with "000000"). Also, responses from international samples in which the question followed the International System of Units have been converted into English units. This led to 6127 valid and interpretable responses for Anchoring 1 - NYC, 6086 valid responses to the Anchoring 2 - Chicago question, 6120 valid responses to the Anchoring 3 -Everest question and 6234 valid responses to the Anchoring 4 -Babies question.

Finally, responses that were above and below -respectively- the high and low anchors have been set to missing. This led to 3691 responses for Anchoring 1 - NYC, 2553 valid responses to the Anchoring 2 - Chicago question, 3364 valid responses to the Anchoring 3 -Everest question and 3983 valid responses to the Anchoring 4 -Babies question. T-tests confirmed that subjects in the low anchor conditions provided lower estimates than subjects in the high anchor condition. Participants in the high anchor group asked to estimate the distance from San Francisco to NYC provided a mean of 3954 miles (SD=1088), whereas participants in the low anchor conditions provided a mean estimate of 2797 miles (SD=873; t(3689)=35.677, p<.001). Participants in the high anchor group asked to estimate the population of Chicago provided a mean estimate of 3,068,940.2 inhabitants (SD=1,230,680.4), whereas participants in the low anchor conditions provided a mean estimate of 1'040'715.5 inhabitants (SD=982,089.6; t(2551)=46.02, d=1.82. Participants in the high anchor group asked to estimate the height of Mt. Everest provided a mean estimate of 34,012.58 feet (SD=9,455.05), whereas participants in the low anchor conditions provided a mean estimate of 11,402.23 feet (SD=10,270.06; t(3362)=65.66, d=2.26. Participants in the high anchor group asked to estimate the number of babies born per day in the US provided a mean estimate of 25,712.59 babies (SD=14,349.93), whereas participants in the low anchor conditions provided a mean estimate of 2,527.82 babies (SD=5341.44; t(3362)=71.67, d=2.47.
Visual inspection of the distributions of the Anchoring dependent variables showed that typical assumptions of t-test were violated.

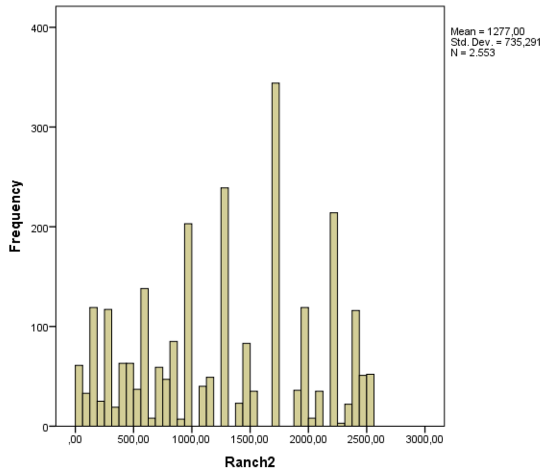


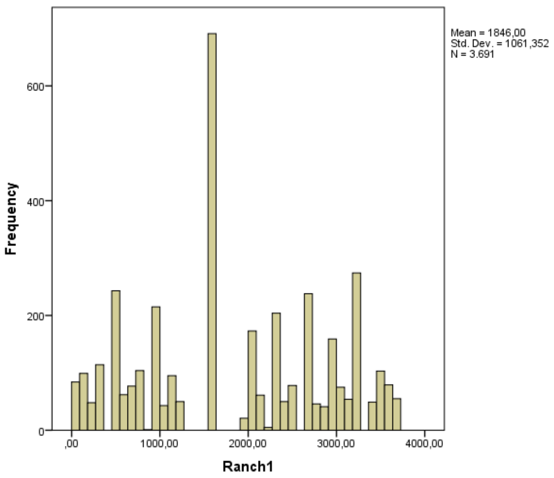


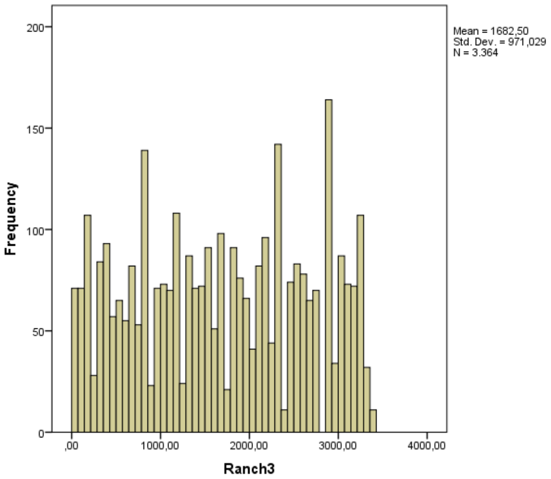


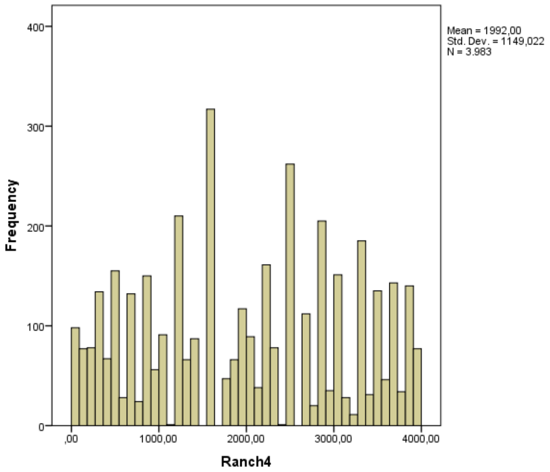


Distributions of raw data are multimodal, asymmetric and negatively skewed. Hence we rank-transformed the distributions in order to meet the assumptions of the t-test. This solution provided slightly more normal distributions. We refrained from further transformations as they would have sharply modified the original data.



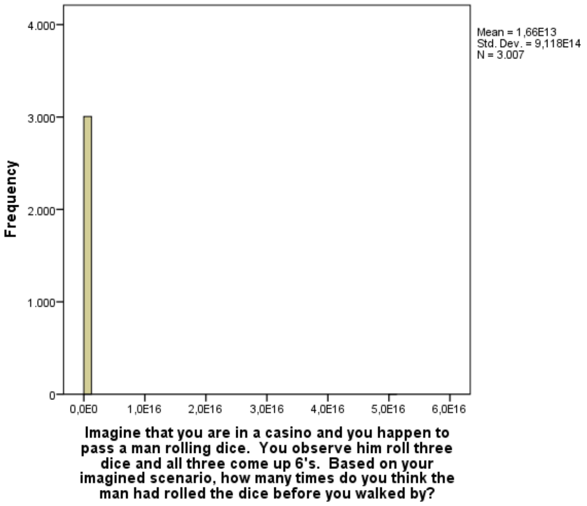


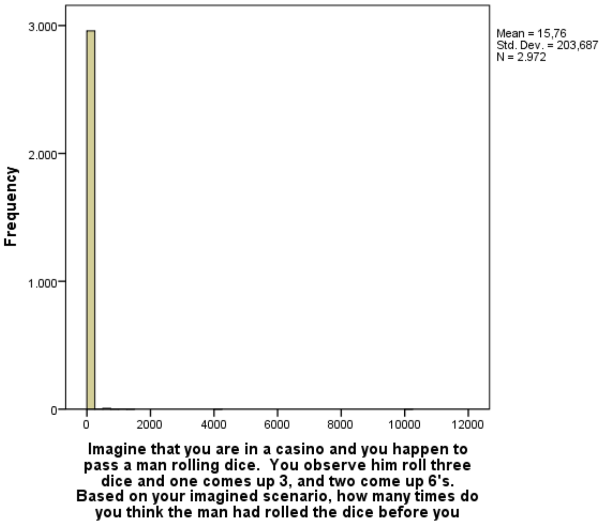




The results of confirmatory analyses on the rank-transformed DVs are the following: Anchoring 1: t(3689)=35.98, d=1.18; Anchoring 2: t(2551)=45.46, d=1.8; Anchoring 3: t(3362)=63.79, d=2.2; Anchoring 4: t(3981)=72.79, d=2.4;

4 Retrospective gambler's fallacy (Oppenheimer & Monin, 2009). In the original study, participants believed that a sequence of previous dice rolls was more than three times as long when a set of three 6's was observed (M = 34.2) than when there were two 6's and a 3 (M = 3.2), t(57) = 2.65, p < .05, Cohen's d = 0.69). In the replicated study, 5979 participants provided valid responses. The mean response when subjects were asked to estimate the number of dice rolls after a set of three 6's was observed was M=1.85*10^13 (SD=9.6*10^14) and M=14.32 (SD=194) when asked to estimate the number of dice rolls after a set of two 6's and a 3 was observed (t(5977) = 1.098, p = .27, Cohen's d = 0.02). Results of the t-test are uninterpretable because the data violate the assumption of normality (see histograms below).





The distribution in this case has been transformed taking the square root of responses and dropping 37 responses that were above 3 SDs from the mean. The histogram of the normalized distribution is presented below and the results of the t-test run on the normalized distribution are presented in Table S1

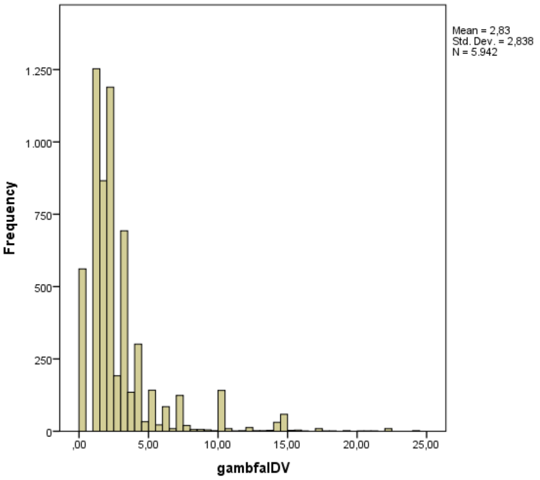


Table S1. Confirmatory t-tests of the Retrospective Gambler's Fallacy after the dependent variable have been normalized.

Group	N	Mean (SD)	t(df)	d
Three 6	2680	3.76 (3.32)	24.01 (5940)	.623
Two 6 one 3	3262	2.06 (2.07)		

5 Low-vs.-high category scales (Schwarz et al., 1985). In the original study, 16.2 percent of the respondents who were presented the low category range reported watching TV for more than two and a half hours, while 37.5 percent of the respondents presented the high category range did so, $\chi^2(1)= 7.7$, $p < .01$. In the replication study, 7.2 percent of the respondents who were presented the low category range reported watching TV for more than two and a half hours, while 24.6 percent of the respondents presented the high category range did so, $\chi^2(1)= 342.39$, $p < .001$, d=.49. No participant's data has been excluded from analysis; 445 subjects failed to provide valid responses.

6 *Norm of reciprocity (Hyman and Sheatsley, 1950).* In the original study, 73% of participants supported allowing communist reporters into America when it was asked after a question about allowing

American reporters into the communist country, but only 36% supported allowing communist reporters into America when the question about it was asked first. In the present replication, 77.3% of participants supported allowing reporters from N. Korea into America when it was asked after a question about allowing American reporters into N. Kore, but only 63.9% supported allowing reporters from N. Korea into America when the question about it was asked first. To test for replication we conducted a chi-square analysis on statement-agreement with condition (US-first, North Korea-first) as a between-subjects variable ($\chi^2(1)= 135.32$, $p < .001$, d=.3). No participant's data has been excluded from analysis; 68 participants failed to provide a valid response.

7 Allowed/Forbidden (Rugg, 1941). In the original study, 62% of participants indicated “No” when asked if speeches against democracy should be allowed, but only 46 percent indicated “Yes” when asked if these speeches should be forbidden. In the replication study, 23.7% of participants indicated “No” when asked if speeches against democracy should be allowed, but only 7.2 percent indicated “Yes” when asked if these speeches should be forbidden. To test for replication we conducted a chi-square analysis on policy choice with condition (allowed, forbidden) as a between-subjects variable ($\chi^2(1)= 3088.69$, $p < .001$, d=1.96). No participant's data has been excluded from analysis; 52 participants failed to provide a valid response.

8 Quote Attribution (Lorge & Curtis, 1936). In the original study, participants indicated more agreement with a quotation from a liked individual than a disliked individual (exact numbers unknown). In the replicated study, participants indicated more agreement with a quotation from G. Washington (M=5.93, SD=2.2) than to Osama Bin Laden (M=5.23, SD=2.11). To test for replication we conducted an independent samples t-test with condition (liked, disliked speaker) as the independent variable and agreement with the quote as the dependent variable (t(6323)=12.79, d=.32). No participant's data has been excluded from analysis; 19 participants failed to provide a valid response.

9 Flag Priming (Carter et al., 2011; Study 2). In the original study, attitudes of participants in the flag-prime condition (M = 3.10) were significantly closer to the Republican end of the scale than were attitudes of participants in the control condition (M = 2.65), t(64) = -2.04, $p < .05$. In the replication study, we used only participants from US samples (N=4979). Also, to ensure participants are examining the photos, those participants who did not provide a “time of day this photo was taken” estimate for all four priming photos have been excluded from analysis (82 participants dropped). One participant failed to provide a valid response to any of the items of the Republican-Democratic scale used as dependent variable (alpha=.68). Attitudes of participants in the flag-prime condition (M = 3.17, SD=1.07) were statistically equal in size than those in the control condition (M=3.15, SD=1.04; t(4894) = .879, p =.38, d=.03.

10 Currency priming and system justification (Caruso et al., 2012). In the original study, participants in the money-prime condition (M = 4.96, SD = 1.27) scored higher on the system justification scale than those in the control condition (M = 3.99, SD = 1.19), t(28) = 2.12, p = .043, d = 0.80. In the replication study, data have been analyzed only from those participants who responded to at least six of the eight system justifications items (11 participants dropped). To test for replication we conducted an independent samples t-test on system justification (alpha=.78) with condition (money-prime, control) as a between-subjects variable (M(prime)=3.58, SD(prime)=.99, M(control)=3.60, SD(control)=.99; t(6331)=-.79, p=.43, d=-.02).

11 *Imagined contact (Husnu & Crisp, 2010, Study 1).* In the original study, participants in the "imagined contact" group scored significantly higher (M = 5.93, SD = 1.67) on the composite measure of intentions to engage in future actual contact than participants in the control group (M = 4.69, SD = 1.26), t(31) = -2.39, p = .023, d = .86. To test for replication we conducted an independent samples t-test on the composite score of the contact intention scale (4 items, alpha=.83) with condition (imagined-contact, control) as the independent variable. The mean of the "Imagined Contact" group at the Contact intentions scale (M=4.83, SD=1.90) is slightly higher than the mean of the Control group (M=4.58, SD=1.98, t(6334)=5.05, p<.001, d=.13). No participant's data have been excluded from analysis; 8 participants did not provide any valid responses at the Intentions of Contact scale.

12 Sex differences in implicit math attitudes and relations with self-reported attitudes (Nosek, Greenwald, & Banaji, 2002). In the original study, women's implicit math attitudes were significantly more negative than men's (Cohen's d = 1.01 in Study 1 and 0.90 in Study 2), and (b) across studies, implicit math attitudes were significantly correlated with explicit math attitudes (r = .42). In the replication study, 4273 participants reported being females and 2060 participants reported being males. 11 participants did not report their gender. Data from the IAT have been analyzed using the D algorithm (Greenwald, Nosek, & Banaji, 2003) with the following features: response latencies < 400ms and >10,000ms have been removed, and trial latencies have been calculated from the beginning of the trial until the time of a correct response. 6185 subjects provided enough valid information to compute an IAT score. 334 participants have been dropped for excess of errors (>40% errors on a single block or > 30% errors overall) or because they did not respond to a particular item for both math and arts, or because they did not respond to at least 6 of the eight total explicit attitude items. To test for replication of (a) we conducted an independent samples t-test on implicit math attitudes with participant sex (men, women) as the independent variable (t(5840)=19.277, p<.001; d=.53). To test for replication of (b) we computed a correlation coefficient for the relationship between implicit and explicit math attitudes (alpha=.89; r=.38, p<.001, d=.79).

Exploratory moderation analyses

To test whether the location of data collection (US vs International samples) and the setting of data collection (Laboratory vs On-line) moderate the magnitude of the effects under investigation we estimated a series of ANOVA models on each effect's dependent variable with 1) experimental conditions, 2) US vs Intl categorical moderator and 3) Lab vs online categorical moderator as between subjects factors, plus the two-way interactions between 3) "Experimental Conditions" and "US vs INTL" moderator and 4) "Experimental Conditions" and "Lab vs Online" moderator. The critical tests of the two interaction effects are summarized in Table S2.

For the "Correlations between Implicit and Explicit attitudes" effect we estimated two hierarchical linear regression (one for each moderator) with the Implicit measure as DV and the explicit measure as IV. Then we added the categorical moderator (dummy coded 0,1) in a second step and the product term of the interaction in a third step. Neither the product term involving the "US vs Intl samples" moderator (Delta

R^2<.001 F(1,5989)=.413, p=.520) or that involving the "Lab vs Online" moderator (Delta R^2<.001 F(1,5989)=2.799, p=.094) change significantly the amount of variance explained by the model in the dependent variable.

The order in which studies have been presented also has been tested as a potential moderator with a series of ANOVA models on each effect's dependent variable with "experimental condition" as a fixed factor and "Order of presentation" (11 levels) as a random factor. No interaction is significant for alpha=.05. For priming studies, an a-priori contrast has been computed comparing the first level (study presented first in the sequence) with all the others. No contrast is significant for alpha=.05.
To test the specific "Flag Priming" moderators we estimated (on US participants only) a series of hierarchical regression models adding the product terms in the last step. Table S2 provides a summary of the results.

Materials and data files

Study materials, analyses scripts and data files can be downloaded at the ManyLabs page of the Open Science Framework: https://opensciencframework.org/project/WX7Ck/
The main data source is the file named "Full.Dataset.De-Identified.sav". In this SPSS data file, variable and value labels work as a codebook to facilitate secondary analyses. The main SPSS syntax file for data preparation and confirmatory tests is named "Syntax.manylabs.sps". This syntax can be run as a whole to generate two data files called "effect.sizes.overall.all" and "effect.sizes.all" that provide effect sizes for each study (the first) disaggregated by sample (the second). Saving the manylabsdata after the syntax has been run as a whole generates the file "ManylabsData.Factors.DVs" that can be used as a data source for moderation analyses. Syntax for moderations analyses are in the files named "ModerationAnalyses(lab&web&flagpriming).sps" and "Order.effects.moderations.sps". A second aggregated data file is named "effectsizes.graphdata1.sav" and it can be used as a source for creating the graphs (syntax: "Graph.sps"). The "syntax.manylabs.sps" code generates exact confidence intervals around the sample and aggregate effect sizes that are based on the central normal distribution. The noncentrality exact confidence intervals around the original effects and around the aggregate replication effect sizes have been computed in R with the package MBESS (Kelley & Lai, 2012) according to the methods described in Kelley (2007a, 2007b). . The R script is named "ESCIOrginalstudies.R". Comments in all syntax files are marked by "*" or "#" and identify the specific study that is being analyzed and the analyses that are carried out by the preceding or following code. Q and I^2 statistics have been computed in R with the package "metafor" (Viechtbauer, 2010). Additional file formats are available in the Datasets.zip archive at the ManyLabs page of the Open Science Framework.

Table S2. Hierarchical Linear Regression Models Testing Flag Priming Moderators

Test of moderator 1: Predictor added	R^2	ΔR^2	ΔF	df1	df2	p
1 - Experimental condition (dummy coded 0,1)	.000	.000	.710	1	4687	.399
2 - How much do you identify with being American?	.050	.050	248.13	1	4686	.000
3- Condition*moderator	.051	.001	.508	1	4685	.476
Test of moderator 2: Predictor added	*R^2*	*ΔR^2*	*ΔF*	*df1*	*df2*	*p*
1 - Experimental condition (dummy coded 0,1)	.000	.000	.528	1	4683	.467
2- To what extent the typical American is a Rep. or Democrat?	.001	.000	2.23	1	4682	.136
3- Condition*moderator	.001	.000	.064	1	4681	.800
Test of moderator 3: Predictor added	*R^2*	*ΔR^2*	*ΔF*	*df1*	*df2*	*p*
1 - Experimental condition (dummy coded 0,1)	.000	.000	.668	1	4680	.414
2 - To what extent the typical American is conservative or liberal?	.007	.007	31.47	1	4679	.000
3- Condition*moderator	.007	.000	.240	1	4678	.624
Test of two-way interaction 1: Predictor added	*R^2*	*ΔR^2*	*ΔF*	*df1*	*df2*	*p*
1 - Experimental condition (dummy coded 0,1)	.000	.000	.64	1	4672	.423
2 - How much do you identify with being American?	.050	.050	244.01	1	4671	.000
3- To what extent the typical American is a Rep. or Democrat?	.051	.001	3.655	1	4670	.056
4- Condition*moderator1*moderator2	.051	.001	3.260	1	4669	.071
Test of two-way interaction 2: Predictor added	*R^2*	*ΔR^2*	*ΔF*	*df1*	*df2*	*p*
1 - Experimental condition (dummy coded 0,1)	.000	.000	.787	1	4670	.375
2 - How much do you identify with being American?	.050	.050	245.08	1	4669	.000
3 - To what extent the typical American is conservative or liberal?	.056	.006	28.06	1	4668	.000
4 - Condition*moderator1*moderator2	.056	.001	3.80	1	4667	.051
Note: In block 0 intercept added in all models						

Appendix A: Recruiting the Collaborative Team

Text of recruitment message posted to the Open Science Framework Google Group (https://groups.google.com/forum/?hl=en&fromgroups=#!forum/openscienceframework).

Subject: Collaborators Needed for "Many Labs" Replication Project

Dear Colleagues,

We are currently seeking collaborators for a wide-scale replication project to be submitted for consideration in a special issue of *Social Psychology* “Replications of Important Results in Social Psychology.”

The goal of the project is to take a small set of documented effects in social psychology that are extremely easy to administer. We will perform replications in as many independent labs as possible. Some of the effects are known to be highly replicable, for others there is less knowledge about replicability. Also, some are thought to depend on the social context or sample, whereas others may not be. This will allow the results to be compared across various locations, participant populations, and lab set-ups, in order to learn more about the role these factors play in replicability.

We have identified an initial list of about 12 effects that take just seconds or a few minutes each to administer. All effects can be automated and will be run through a single experiment script (using the web-based Project Implicit infrastructure). This way, all replication teams can run the identical study script in their laboratories with their own samples, and it will only take a few clicks to launch.

Participants will be able to complete the study in 10-15 minutes. This way, the study could be administered as an independent session or in “extra time” after another data collection. We are also happy to work with you to resolve any issues that may impede your ability to administer the experiment.

With efforts such as the Reproducibility Project already underway, understanding the situational variables that influence replication takes on an important role. In addition to aiding in the advancement of this research, all participating researchers will be credited as co-authors on the final publication.

If you are interested in joining the project or would like to get further information or ask questions, please contact the investigators at manylabsproject@gmail.com.

Sincerely,
The Many Labs Team
Richard A. Klein, University of Florida
Dr. Kate A. Ratliff, University of Florida
Dr. Brian A. Nosek, University of Virginia

Appendix B: Contextual Information about Sample and Setting

List of contextual information to be collected from researchers.

1. Features of the experimenter(s) – Gender, ethnicity.
2. How many participants were administered the survey at once?
3. Were participants separated or not?
4. Did participants participate in another study prior to being administered the replication? If so, please indicate what that study was.
5. How were participants recruited?
6. How were participants compensated (paid/volunteer/course credit)?
7. Was the study presented in the original language?

Appendix C: Demographics

List of demographic questions to collect from participants:

1. Age
2. Race and ethnicity
3. Political ideology
4. Country of citizenship
5. Native language
6. Major (if any)
7. Prior exposure to included experiments (we will provide a list with brief descriptions, and they will mark if they have seen or learned about it previously). *

*Only collected for the MTurk and Project Implicit samples due to time concerns

Appendix D: Flag Priming Moderator Items

Additional items for flag priming study (administered at the end of the study package).

1. How much do you identify with being American? (1, not at all – 11, very much)*
2. To what extent do you think the typical American is a Republican or Democrat? (1, Democrat – 7, Republican)
3. To what extent do you think the typical American is conservative or liberal? (1, Liberal – 7, Conservative)**

* For international replications, this question will be adapted to "How much do you identify with Americans?"
** For international replications, "liberal" may be replaced with "left" and "conservative" may be replaced with "right", as deemed appropriate by the local researcher.

References

Kelley, K., & Lai, K. (2012). MBESS: Methods for the Behavioral, Educational, and Social Sciences. R package version 3.3.3. http://CRAN.R-project.org/package=MBESS

Kelley, K. (2007a). Methods for the behavioral, educational, and social sciences: an R package. *Behavior Research Methods*, *39*(4), 979–84.

Kelley, K. (2007b). Confidence Intervals for Standardized Effect Sizes: Theory, Application, and Implementation. *Journal of Statistical Software*, *20*(8), 1-24.

Wolfgang Viechtbauer (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36(3)*, 1-48. URL http://www.jstatsoft.org/v36/i03/.