Investigating the replicability of the social and behavioural sciences

Investigating the replicability of the social and behavioral sciences

Andrew H. Tyner, Anna Lou Abatayo, Mason Daley, Samuel Field, Nicholas Fox, Noah A. Haber, Krystal M. Hahn, Melissa Kline Struhl, Brinna Mawhinney, Olivia Miske, Priya Silverstein, Courtney K. Soderberg, Theresa Stankov, Ahmed Abbasi, Christopher L. Aberson, Balazs Aczel, Matúš Adamkovič, Nihan Albayrak, Peter J Allen, Michael Andreychik, Eli Awtrey, Erick Axxe, Flavio Azevedo, Miles D. Bader, Bence Bago, James Bailey, Marjan Bakker, Gabriel Banik, George C. Banks, Ernest Baskin, Anatolia Batruch, Annika Beatteay, Sophie M. Behr, Nicholas Berente, Zachariah Berry, Jędrzej Białkowski, Bojana Bodroža, Laura Boeschoten, Miklos Bognar, Christian Bokhove, Diane Bonfiglio, Robin Bouwman, Timothy F. Brady, Scott Braithwaite, Gabriel Briceño Jiménez, Cameron Brick, Traci Bricka, Roman Briker, Annette N. Brown, Gordon D A Brown, Robbie C.M. van Aert, Kathryn Caldwell, Sara Capitan, Tabaré Capitán, Jesse Chandler, Tessa Charles, Christopher R. Chartier, Rahul Chawdhary, Kent Jason Cheng, William J. Chopik, Bruce Clark, Victoria E. Colvin, C. Cozette Comer, Giulio Costantini, Tom Coupé, Jamie Cummins, Aneta Czernatowicz-Kukuczka, Joshua de Leeuw, David Dobolyi, James N. Druckman, Jianhua Duan, Marin Dujmović, Daniel J. Dunleavy, Patrick K. Durkee, Cécile Emery, Kevin M. Esterling, Thomas R. Evans, Anna Fedor, Belén Fernández-Castilla, Nathan Fiala, James G. Field, Nathan Fong, Miguel A. Fonseca, Alexandra L.J. Freeman, Jeremy Freese, Sandra J. Geiger, Jing Geng, Laura M. Getz, Linda Marjoleine Geven, Ilka Helene Gleibs, Donna Pamella Gonzales, Janaki Gooty, Amélie Gourdon-Kanhukamwe, Cristina Greculescu, Siobhán M. Griffin, Lusine Grigoryan, Martina Grunow, Nicholas Gunby, Braeden Hall, Paul H. P. Hanel, Erin E. Hannon, Sam Harper, Marco Jürgen Held, Louis Hickman, Nathan C. Higgins, Svenja Hippel, Sven Hoeppner, Sanghyun Hong, Thomas J. Hostler, Michael Inzlicht, Kamil Izydorczak, Bastian Jaeger, Kristin Jankowsky, Johannes Jarke-Neuert, Matthew Jensen, Biljana Jokić, Daniel Jolles, Phillip Jolly, Angela M. Jones, Marie Juanchich, Pavol Kačmár, Hansika Kapoor, Andjela Keljanovic, Samjhana Koirala, Marta Kołczyńska, Dimitra Kouroupaki, Ulrich Kühnen, Michelangelo Landgrave, Michael J. Larson, Lyonel Laulié, Alice C E Lawrence, Joel M. Le Forestier, Katelin E. Leahy, Sungmok Lee, Jared Leslie, Savannah C. Lewis, Christopher Limnios, Hause Lin, An-Chiao Liu, John Wills Lloyd, Elliot A Ludvig, Dermot Lynott, Jordan MacDonald, Peter Mallik, Daniel J. Mallinson, Daniele Marinazzo, Corinna S. Martarelli, Joshua Matacotta, Andrew McBride, Cillian McHugh, Gail McMillan, Esteban Méndez, Mitchell Metzger, Michalis P. Michaelides, Johannes Michalak, Leticia Micheli, Jeremy K. Miller, Marina Milyavskaya, Daniel C. Molden, Ambar G. Monjaras, David Moreau, Audrey Morrow, Cristóbal Moya, Liad Mudrik, Laetitia B. Mulder, Katie A. Munt, Arijit Nandi, Kathryn Nason, Carolin Nast, Gideon Nave, Heinrich H. Nax, Florian Neubauer, Phuong Linh L. Nguyen, Austin Lee Nichols, Gustav Nilsonne, Ernest O'Boyle, Jule Oettinghaus, Jeewon Oh, Adoril Oshana, Thomas Ostermann, Rachel P. Ostrowski, Abiola Oyebanjo, Radoslaw Panczak, Jamie Patrianakos, Ignacio Pavez, Yuri G. Pavlov, Sofia Persson, Marco Perugini, Kim Peters, Constant Pieters, Vladimir Ponizovskiy, Nathaniel D. Porter, Jason M. Prenoveau, Danka Purić, Mariah F. Purol, Arathy Puthillam, Kimberly A. Quinn, Marco Ramljak, W. Robert Reed, Michaela Ritchie, Margaret Ritzau, Sean Patrick Roche, Romina Rodela, Jan Philipp Röer, Ivan Ropovik, Jacob Rothschild, Justine Saal, Hani Safadi, Jason Samaha, Mary Sanchez, Soorya Sankaran, David Santos, Amanda C. Sargent, Marian Sauter, Kathleen Schmidt, Landon Schnabel, Amber N Schroeder, Sebastian W. Schuetz, Brendan A. Schuetze, Michael Schulte-Mecklenbeck, Astrid Schütz, Eric L. Sevigny, Ellie Shackleton, Richard M. Shafranek, Samuel Shaki, Shishir Shakya, Miroslav Sirota, Matthew Ryan Sisco, Maksim M. Sitnikov, L. Robert Slevc, Laura Smalarz, Colin Tucker Smith, Joel S. Snyder, Nicolas Sommet, Fatih Sonmez, Barbara A. Spellman, Natalia Stanulewicz-Buckley, George Stock, Chris N. H. Street, Eirik Strømland, Tina Sundelin, Moin Syed, Anna Szabelska, Barnabas Szaszi, Ewa Szumowska, Anirudh Tagat, Susanne Täuber, Louis Tay, Stuti Thapa, Jason Thatcher, Domna Tsaklakidou, Lars Tummers, Elise Turkovich, Melba Verra Tutor, Karolina Urbanska, Anna Elisabeth van 't Veer, Marcel van Assen, Niels van de Ven, Elisabeth Julie Vargo, Leigh Ann Vaughn, Simine Vazire, Jentien M. Vermeulen, Diem Thi Hong Vo, Victor Volkman, Eric-Jan Wagenmakers, Deliah Wagner, Lukasz Walasek, Frank Walter, Lara Warmelink, Liuqing Wei, Marie Isabelle Weißflog, Nicholas Weller, Aaron L. Wichman, Jonathan Wilbiks, Jamal R. Williams, Kelly Wolfe, Finnian Wort, Ryan Wright, Jesper N. Wulff, Xindong Xue, Veronica X. Yan, Yuzhi Yang, Sangsuk Yoon, Iris Žeželj, Yinxian Zhang, Ignazio Ziano, Cristina Zogmaister, Zorana Zupan, Rolf A. Zwaan, Brian A. Nosek, & Timothy M. Errington

Abstract

We attempted replications of 274 claims of positive results from 164 quantitative papers published from 2009 to 2018 in 54 journals in the social and behavioral sciences. Replications were high-powered on average to detect the original effect size (Median = 99.6%), used original materials when relevant and available, and were peer-reviewed in advance through a standardized internal protocol. Replications showed statistically significant results in the same pattern as the original study for 151 of 274 claims (55.1% [95% CI 49.2 - 60.9%]) and for 80.8 of 164 papers (49.3% [95% CI 43.8 - 54.7%]) weighed for replicating multiple claims per paper. We observed modest variation in replication rates across disciplines (Range 42.5% to 63.1%) though some estimates had high uncertainty due to small sample size. For claims where effect sizes could be converted to Pearson's *r*, the median effect size was 0.25 [95% CI 0.21 - 0.27] for original studies and 0.10 [95% CI 0.09 - 0.13] for replication studies, a 58.1% [95% CI 44.2 - 65.0%] reduction in correlation and a 82.4% [95% CI 67.8 - 88.2%] reduction in shared variance. Thirteen methods for evaluating replication success provided estimates ranging from 28.6% to 74.8% (median = 49.3%), though most methods could only be applied to a subset of the replications conducted (median = 92.1%; range 63.5% to 100.0%). Some decline in effect size and significance is expected based on power to detect original effects and regression to the mean due to replicating only positive results. The conditions that promote or inhibit replicability are worthy of additional investigation.

Keywords: replication, credibility, reliability, validity, economics, political science, psychology, marketing, sociology, finance, management, public administration, organizational behavior, education, criminology, health research

A central aim of science is to discover regularities in nature. If a claimed discovery is true, independent researchers should be able to conduct a similar investigation and reach similar conclusions. A replication attempt is testing the same research question as a prior investigation with independent evidence, whether the evidence is a new data collection or existing secondary data that was not used in the prior investigation.[1,2]

Across several prior replication studies in the social and behavioral sciences totaling hundreds of replication attempts, approximately half of well-powered replication studies provided statistically significant evidence in the same direction as an original finding.[3–11] Moreover, observed effect sizes in replication studies were about half as large as effect sizes in original studies on average.[10] Similar results have been observed in other fields, such as preclinical cancer biology.[12]

Popular explanations for the low observed replication success rates include underreporting of negative or inconclusive evidence for claims; high sampling error from small samples and measurement error due to unreliable measures, coupled with a statistical threshold ($p < .05$) that serves as a publication filter; low rigor and quality control in research design and measurement; and questionable research practices that inflate the likelihood of obtaining positive outcomes.[10,13–20] These factors are compounded by a research culture that rewards novel and "interesting" findings and discourages error correction.[21–23] Replication failures are not necessarily due to the original findings having low credibility. Low replication rates can also be due to false negatives, poorly designed replications, selecting only positive results for replication, and differences between original and replication studies that are initially perceived as unimportant.[3,24] More broadly, there are conceptual challenges in deciding what is a replication of a prior finding and how to assess whether a replication attempt succeeded. Attempting replications provides a grounded context to wrestle with those challenges.

Identifying the potential causes of replication failures assumes that the existing evidence of replication failures is itself replicable. Most evidence comes from studies that examine replication outcomes within a single discipline and with relatively small samples. Here, we report a systematic replication of quantitative published claims conducted as part of the Defense Advanced Research Projects Agency (DARPA) Systematizing Confidence in Open Research and Evidence (SCORE) program. Papers and claims were drawn from a sample of well-known journals selected to represent a diversity of subdisciplines across the social and behavioral sciences (Table S1). The selected journals are aggregated for expository purposes into six disciplines: 11 journals in Business (including Organizational Behavior, Management, and Marketing), 9 journals in Economics (including Finance), 7 journals in Education, 7 journals in Political Science (including Public Administration), 13 journals in Psychology (including Health), and 7 journals in Sociology (including Criminology). See Supporting Information for outcomes separately by subdiscipline, selection effects, effect size comparisons, outcomes across claims, and subset analyses (Tables S8-S18; Figures S10-14). Papers and claims eligible for inclusion had to be quantitative research of any kind and contain a statistical inference that identified a positive result using non-simulated human data including any level of human organization (e.g., individuals, families, political entities, firms, economic units).

Results

We first summarize how the sample of completed replications compares with the sample of papers and claims. Then, we assess the replication outcomes across a variety of criteria[12]. Primary reporting emphasizes the two most reported replication metrics: statistical significance and effect size comparisons. This is followed by summarizing the same evidence with 12

additional metrics that have been used to evaluate replication success, each with different assumptions and limitations.

Replications completed by discipline and year in comparison with the sample of papers and claims

We randomly selected papers and claims from those published within the selected journals and time frame (see Methods). However, replication attempts were constrained by feasibility, access to resources, and availability of researchers with relevant interest, expertise, and instrumentation, leading to potential selection effects.

Table 1 illustrates the proportion of the selected papers by discipline for which replication attempts were started and completed. Representativeness across disciplines was maintained during identification and extraction of claims because we used random sampling strategies. Most of the change in representativeness occurred due to the non-random process of selecting and starting a replication study: education and political science decreased in relative proportion of the sample, and psychology increased. Compared with the original sample of papers (first row), the proportion of completed attempts of unique claims (last row) was within 3% for economics, education, political science, and sociology, and more notable variation was observed for business and psychology. Representativeness of replication attempts was relatively steady by year compared to the sample of papers as reported in the supporting information.

Table 1. Papers and claims selected, and replication attempts, by discipline.

	Business	Economics	Education	Political Science	Psychology	Sociology	Total
				n (%)			
Claims selected							
Papers with claims	766 (19.6%)	673 (17.3%)	445 (11.4%)	551 (14.1%)	950 (24.4%)	515 (13.2%)	3900 (100%)
Papers eligible for replication	294 (19.6%)	255 (17.0%)	172 (11.5%)	212 (14.1%)	369 (24.6%)	198 (13.2%)	1500 (100%)
Papers with multiple claims	38 (19.0%)	33 (16.5%)	23 (11.5%)	32 (16.0%)	49 (24.5%)	25 (12.5%)	200 (100%)
Papers with single claim	256 (19.7%)	222 (17.1%)	149 (11.5%)	180 (13.8%)	320 (24.6%)	173 (13.3%)	1300 (100%)
Replications attempted							
Papers with replication started	46 (23.2%)	27 (13.6%)	14 (7.1%)	18 (9.1%)	65 (32.8%)	28 (14.1%)	198 (100%)
Papers with replication attempts completed	36 (22.0%)	24 (14.6%)	13 (7.9%)	15 (9.1%)	58 (35.4%)	18 (11.0%)	164 (100%)
Total replication attempts of claims	42 (14.2%)	40 (13.5%)	28 (9.5%)	45 (15.2%)	108 (36.5%)	33 (11.1%)	296 (100%)
Unique claims with replication attempts	36 (13.1%)	38 (13.9%)	28 (10.2%)	45 (16.4%)	94 (34.3%)	33 (12.0%)	274 (100%)

Note: The row "Total replication attempts of claims" is a count of all replication attempts with recognition that some claims were replicated multiple times (see Methods). The row "Unique claims with replication attempts" is a count of how many claims had a replication attempt.

We selected papers and attempted replications in two phases (see Methods) with 139 of the completed replications occurring from papers selected during the first phase, and 25 from papers selected during the second phase. Replication success rates were similar between the first (49.5% statistically significant with the same pattern) and second phase (48.0%).

Evaluating the replication outcome against the null hypothesis of no effect

A common method for evaluating whether a replication supports an original claim is to assess whether the observed replication test statistic meets a statistical threshold (e.g., alpha = .05) with the effect showing the same pattern as the original evidence. This approach is simple to explain and can be applied to a variety of statistical models. Yet, it also has some drawbacks such as binary assessment and failure to incorporate indicators of precision.[8,9,25–27]

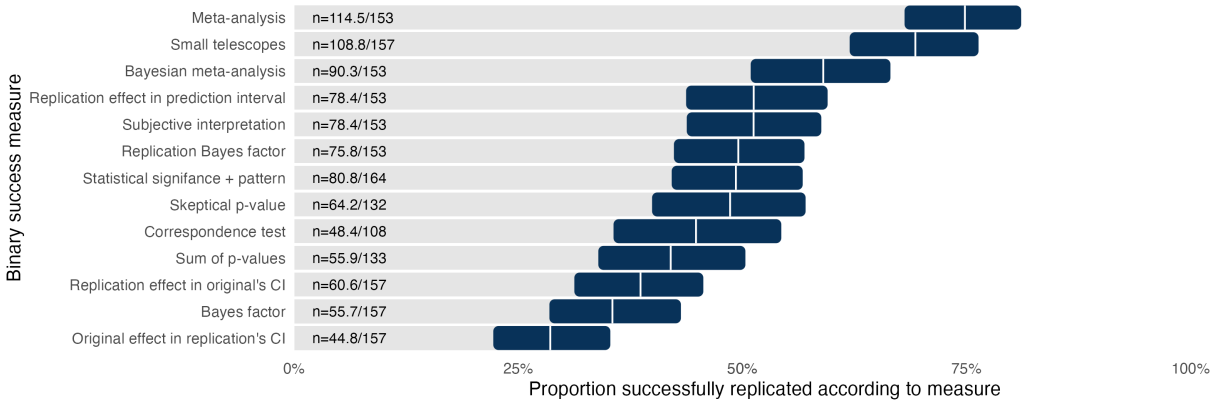
For most papers, we attempted to replicate a single claim; for some papers, we attempted to replicate multiple claims. If the replication success of claims is dependent (perfectly correlated within study), then claims should be weighted so that each study counts as a single observation. If replication successes are independent, then each claim should be counted as a single observation. Weighting by paper, 80.8 of 164 replicated papers had statistically significant findings with the same pattern as the original finding (49.3% [95% CI 43.8 - 54.7%]), 16.0 had statistically significant findings with an opposing pattern (9.7% [95% CI 6.5 - 13.0%]), and 66.2 replications showed a null effect (40.4% [95% CI 34.9 - 45.8%]). Unweighted (by claim), 151 of 274 replicated claims had statistically significant findings with the same pattern (55.1% [95% CI 49.2 - 60.9%]), 24 had statistically significant findings with an opposing pattern (8.8% [95% CI 6.0 - 12.7%]), and 98 showed a null effect (35.8% [95% CI 30.3 - 41.6%]).

We only selected claims for replication that were identified as positive results, usually by exceeding a statistical threshold ($p < .05$). This tends to bias the sample against studies that underestimate effect sizes. We can estimate this expected regression to the mean by estimating the proportion of significant results that would have been selected without the requirement for statistically significant results. For a subsample of 200 papers, using the same process for identifying claims, we selected all outcomes from the paper, regardless of whether they were significant results.[28] We estimated that 2747 of the 3066 claims had significant results (89.6%). The high proportion of significant findings replicates a well-documented bias in the published literature favoring significance.[18,19,29,30] Nevertheless, because the significance rate was not 100%, some regression to the mean is expected in our replication success rates.

Replication attempts can also fail to produce a significant effect due to limited statistical power. Statistical power varies depending on research designs and assumptions. Given the diversity of methods across studies, we calculated power for studies using two different approaches. Under the first approach that was better suited for the original findings with standard effect sizes, the median power to detect the original effect size was approximately 99.6% (alpha = .05): 90.2% of replications had at least 50% power to detect the original effect size, and 87.2% of replications had at least 75% power to detect the original effect size. And, under the approach used when a standard effect size could not be reliably approximated, the median power to detect the original effect size was 99.1% (alpha = .05): 95.1% of replications had at least 50% power to detect the original effect size, and 84.5% of replications had at least 75% power to detect the original effect size. Under a combined approach where each finding draws on the power approach most appropriate to it, the median power to detect the original effect size was 99.6% (alpha = .05):

94.2% of replications had at least 50% power to detect the original effect size, and 86.0% of replications had at least 75% power to detect the original effect size.

Figure 1. Replication success rates across 13 binary assessments for papers. The vertical white line in each row is the estimate of the percentage of papers replicated successfully, and the 95% confidence interval around the estimate is represented by the dark bar. The sample sizes are the weighted number of papers with successful replications first and the number of papers to which that binary assessment could be applied second. See Methods for explanations of each binary assessment. CI = confidence interval.



Evaluating replication success with a variety of binary assessments

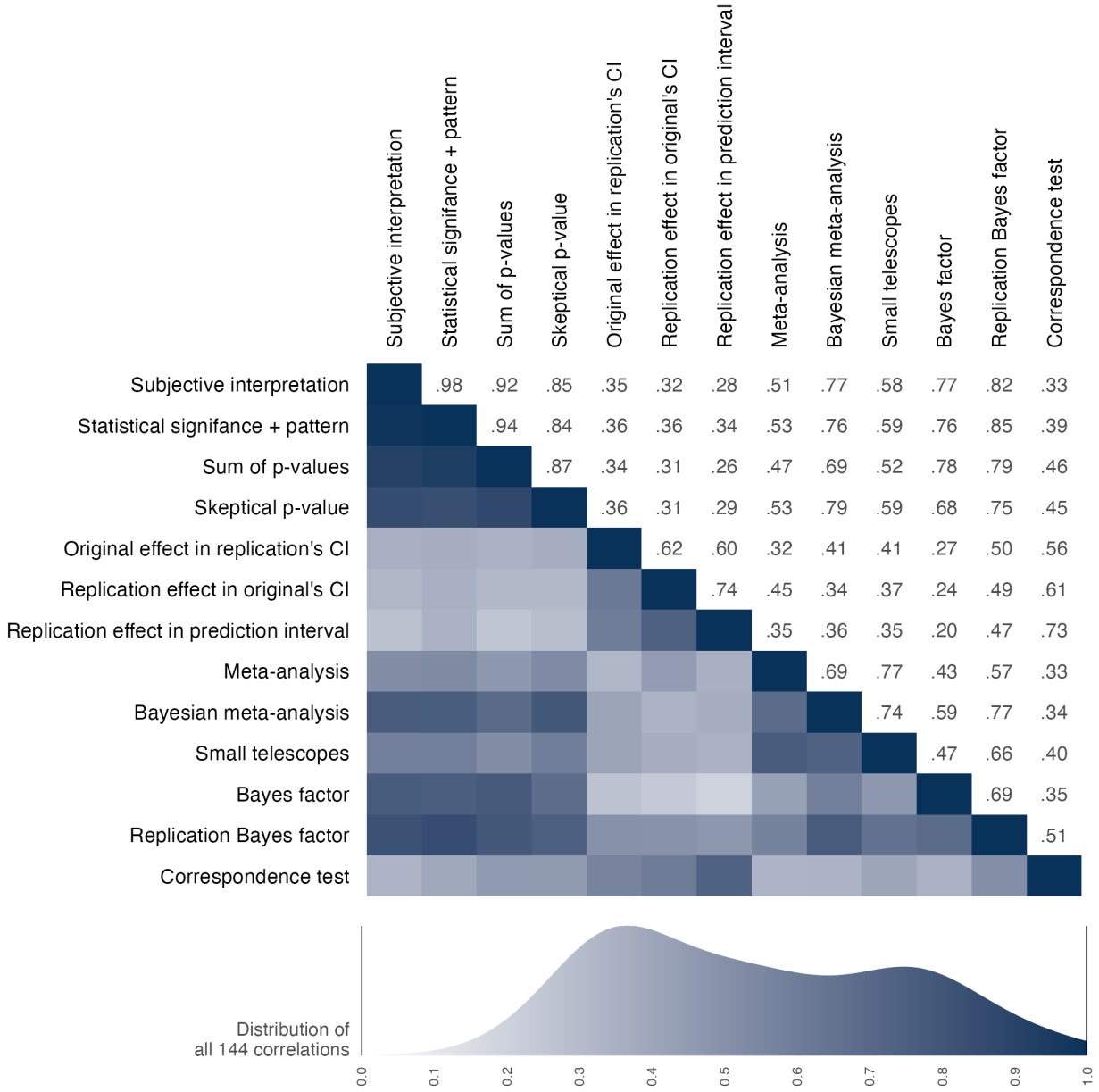
Several methods have been proposed to assess replication success, each with advantages and disadvantages.[31,32] This project highlights a challenging problem: replication success metrics can only be applied to methods that meet their assumptions. In Figure 1, we present 13 replication success metrics along with the number of papers to which that metric could be applied. Some are necessarily binary assessments of replication success or failure, and others were simplified to provide a binary assessment for comparison. The binary assessments were usable for 65.9% to 100.0% of the total sample papers and 55.8% to 100.0% of the total sample of claims. Only the statistical significance metric was applied in all cases. For some of those assessments, statistical assumptions needed to be applied to a subset of original and replication estimates.

Across metrics, replication success rates ranged from 28.6% to 74.8% with a median of 49.3%. The observed variation highlights the impact of the different assumptions underlying each approach. For example, the highest estimate of 74.8% for the meta-analytic combination of the original and replication evidence provides strong evidence of success because it includes original studies that provided evidence of success, and these tests are not independent of the original evidence.

Variation in the observed replication rates are affected by both the assumptions of the binary assessment and by the subsample to which the metric could be used. Figure 2 presents correlations between each pair of binary assessments for replication outcomes to which both methods could be applied. Spearman correlations were positive and relatively high with some exceptions, median = 0.51, range 0.20 to 0.98. For example, analysts almost always interpreted success based on statistical significance (as reflected by a correlation of 0.98), whereas measures that used confidence or prediction intervals showed relatively strong correlations with each other (0.60, 0.62, 0.74) and weaker relations with other measures partly because they

treated replication outcomes both smaller and larger than original outcomes as failures to replicate (Median = 0.35, range = 0.20 to 0.73).

Figure 2: Correlation matrix among binary assessments of replication success across papers. Correlation values are right of the diagonal, and correlation magnitude is visualized left of the diagonal with darker shading indicating stronger correlations. CI = confidence interval.



Evaluating replication effect size against original effect size

Replication can also be defined as the correspondence between effect size coefficients observed in original studies and in their replications. Regardless of whether original studies and replication studies fall on the same side of a statistical threshold, they should produce effects of about the same magnitude. For the purposes of this project, we sought effect size metrics that would be as standardized as possible.

Figure 3. Scatterplot of Pearson's r effect sizes for original and replication studies. Each data point represents the estimated original and replication effect sizes for replicated claims. Size of the bullet is proportional to the number of claims there are per paper to illustrate paper weighting. Replication effect sizes are positive if the observed relationship has the same pattern as the original effect size, and negative if the observed relationship has a different pattern. Data points are classified as successful for effect sizes that achieved statistical significance (p < .05) with the same pattern as the original study and failed for effect sizes that did not. (cf Figure 3 in Open Science Collaboration [2015])[3]

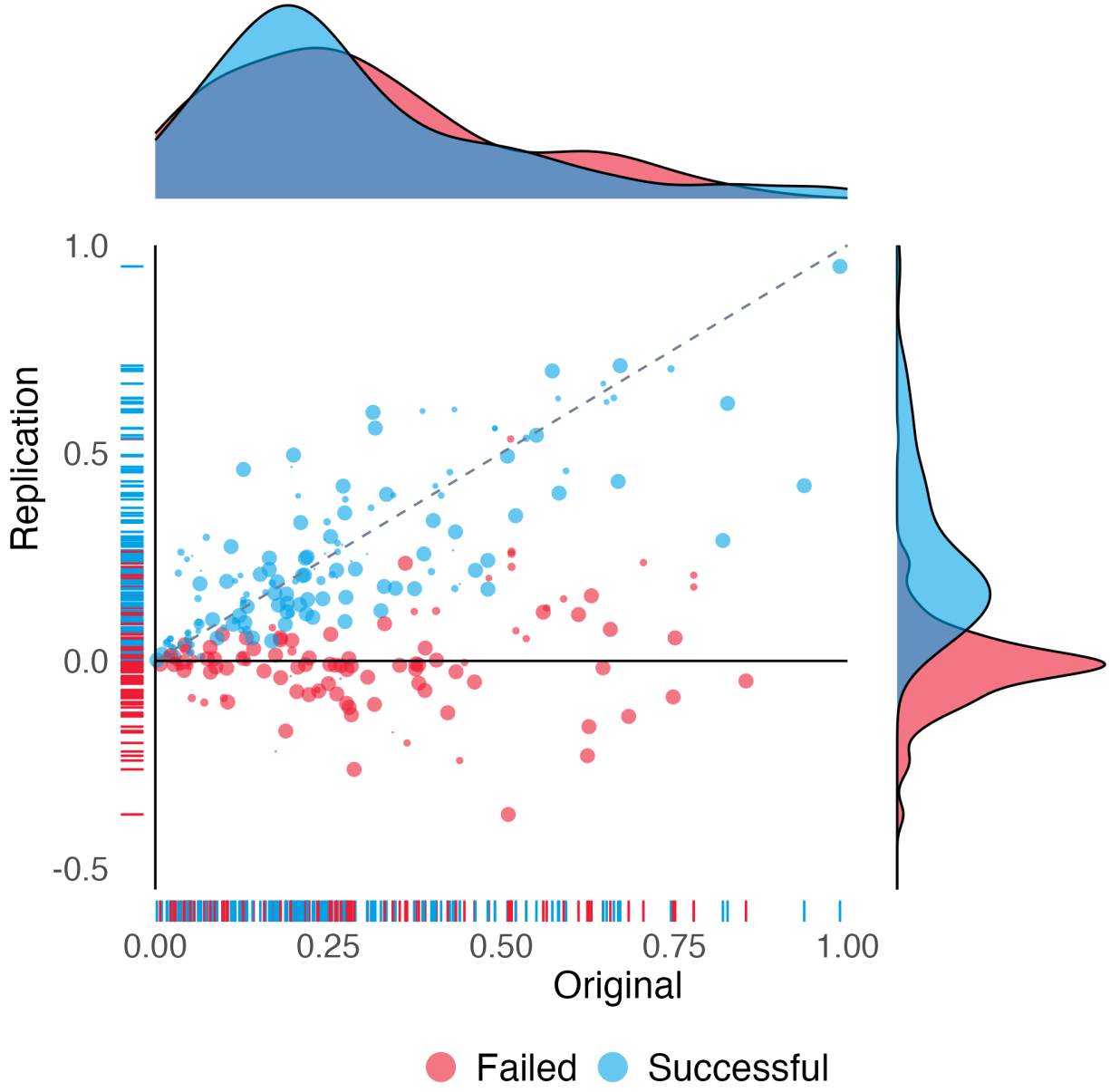


Figure 3 represents the replication effect sizes plotted against the original effect sizes for those claims that could be computed using the Pearson's *r* effect size. Original and replication effect sizes were positively correlated (Spearman's correlation 0.43). The median effect size was 0.25 [95% CI 0.21 - 0.27] for original studies and 0.10 [95% CI 0.09 - 0.13] for replication studies, a

58.1% [95% CI 44.2 - 65.0%] reduction in correlation and a 82.4% [95% CI 67.8 - 88.2%] reduction in shared variance. Data points below the diagonal line are cases in which the replication effect size was smaller than the original effect size. Blue data points were statistically significant replications with the same pattern as the original; red data points were not statistically significant or, if they are below 0, show a different pattern than the original. 126.0 of 157 papers (80.3%) had a smaller effect size for the replications compared with the original studies, and 175 of 249 claims (70.3%) had a smaller effect size for the replications compared with the original studies. Table 2 provides summary statistics of effect sizes by papers and by claims.

Table 2. Original and replication findings by Pearson's *r* effect size by papers and claims.

	Papers (weighted)		Claims (unweighted)	
	Original	Replication	Original	Replication
Number of *r* effect sizes	157		249	
Median [IQR] sample size	206 [125.0]	545 [347.0]	236 [3339.2]	556 [3016.0]
Median Pearson's *r* effect size (SD)	0.25 (0.21)	0.10 (0.17)	0.24 (0.21)	0.13 (0.17)

Note: Sample for this table is the studies for which a Pearson's r could be calculated for the original and replication outcomes. Papers are weighted combinations of claims accounting for multiple claims per paper replicated in some cases. SD=standard deviation, IQR=interquartile range.

Original and replication effects by discipline and year

Table 3 summarizes original and replication outcomes separately by discipline for statistical significance and effect size. By discipline, based on statistical significance, successful replication rates ranged from 42.5% to 63.1% weighted across papers (median 50.0%; chi-square *p*-value = 0.76). A range of 45.5% to 71.4% was observed unweighted across claims (median 50.8%; chi-square *p*-value = 0.13). A similar analysis, reported in the SI, examined variation of replication success by publication year across papers and did not show a significant effect by year ($p = 0.14$).

Table 3. Original and replication outcomes by statistical significance and pattern and by Pearson's *r* effect size for 6 disciplines

	Replication attempt statistical significance and pattern		Median Pearson's *r* effect size (SD)	
	Counts	Percentage	Original estimate	Replication estimate
Business	17.0 / 36	47.2%	0.24 (0.12)	0.10 (0.13)
Economics	10.2 / 24	42.5%	0.28 (0.24)	0.13 (0.20)
Education	8.2 / 13	63.1%	0.15 (0.28)	0.11 (0.10)
Political Science	7.8 / 15	52.0%	0.16 (0.22)	0.05 (0.16)
Psychology	28.4 / 58	49.0%	0.29 (0.22)	0.11 (0.20)
Sociology	9.2 / 18	51.1%	0.10 (0.16)	0.03 (0.17)

Note: Papers are weighted combinations of claims accounting for multiple claims per paper replicated in some cases. Left column indicates the number of successful replications by the total number of papers with replication attempts. Samples for the right columns are the papers for which a Pearson's r could be calculated for original and replication outcomes.

Original and replication effects by new data or secondary data replication

Some replication attempts involved collecting new data and others involved finding secondary data that was not used in the original research (Table 4). Replication attempts using new data were 0.930 [95% CI 0.651 - 1.338] times as likely as those using secondary data to have outcomes that were statistically significant and with the same pattern (Unweighted by claims were 0.986 [95% CI 0.757 - 1.253] times as likely), suggesting similar replicability.

For outcomes that could be estimated with a Pearson's *r* effect size, replication attempts using new data produced effects that were less than half the size of their original effects across papers, and about half across claims. Replication attempts using secondary data showed less decline than those using new data, but from original findings that had smaller effect sizes on average.

The effect size comparisons imply higher replicability for secondary versus new data replication attempts, in contrast to similar replicability observed on the statistical significance metric (Table 4). That could occur if the power of secondary data replication attempts was weaker. Median power estimates across papers are slightly consistent with this possibility (secondary data 99.0%; new data 99.7%). Mean power estimates are more consistent because of a few more weakly powered secondary data replications (secondary data 86.8%; new data 97.3%), and by comparing median power to detect 75% of the original effect size (Table 5).

Table 4. Original and replication outcomes by statistical significance and effect size by new or secondary data replications

	Papers		Claims	
	Original outcome	Replication outcome	Original outcome	Replication outcome
Statistical significance and same pattern				
New data replications	98 / 98 (100.0%)	46.9 / 98 (47.8%)	128 / 128 (100.0%)	70 / 128 (54.7%)
Secondary data replications	66 / 66 (100.0%)	33.9 / 66 (51.4%)	146 / 146 (100.0%)	81 / 146 (55.5%)
Median Pearson's *r* effect size (SD)				
New data replications	0.28 (0.18)	0.11 (0.19)	0.27 (0.18)	0.15 (0.19)
Secondary data replications	0.13 (0.23)	0.10 (0.14)	0.13 (0.23)	0.10 (0.15)

Note: Original outcome refers to the published finding that was the target of the replication study. Replication outcome refers to the results of the replication attempt. New data replications are those that required data collection. Secondary data replications are those that used existing data that was independent of the original investigation. Papers are weighted combinations of claims accounting for multiple claims replicated in some papers. SD=standard deviation.

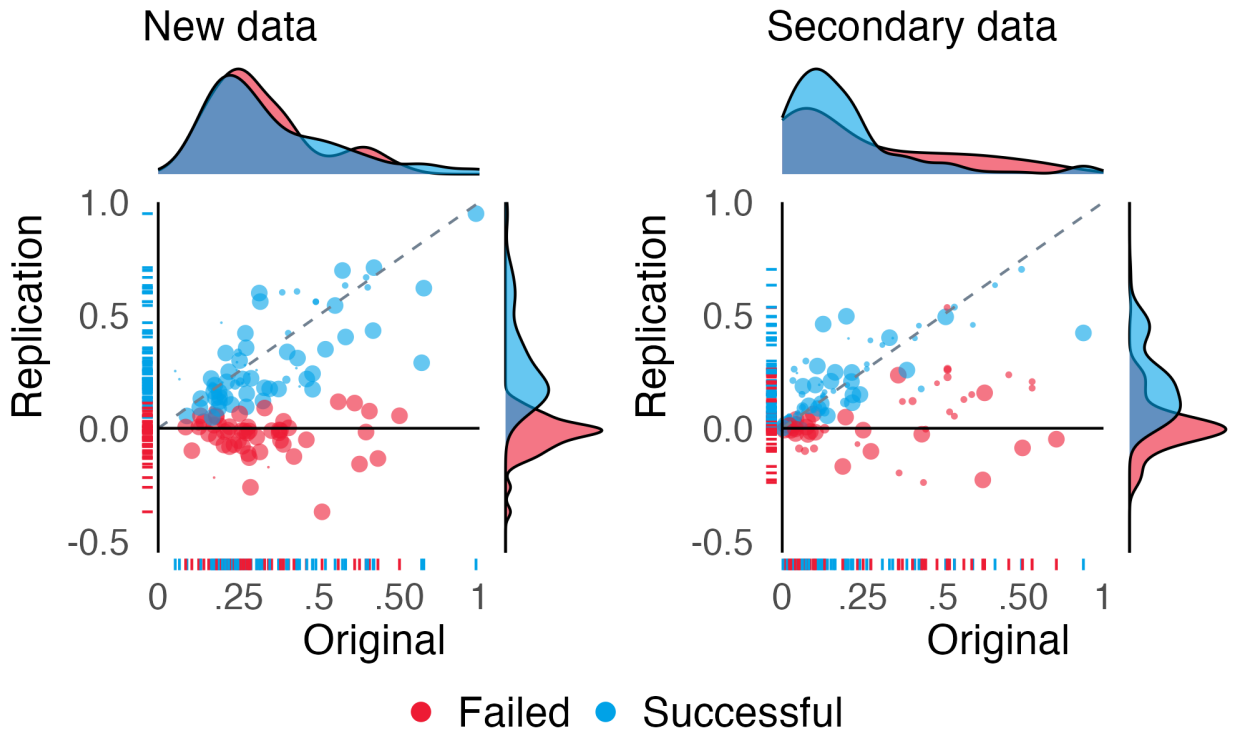
Table 5. Median power to detect 75% of the original effect size

	Papers	Claims
New data replications	94.5%	94.1%
Secondary data replications	83.7%	89.7%

Note: Original effect size refers to the published finding that was the target of the replication study. New data replications are those that require data collection. Secondary data replications are those that used existing data that was independent of the original investigation.

Figure 4 reproduces the scatterplots comparing original and replication effect sizes separately for new data (left panel) and secondary data (right panel). Among effect sizes that could be converted to Pearson's r, 86.8% (83.3 of 96) of new data replication attempts, and 70.0% (42.7 of 61) of secondary data replication attempts, had a weaker effect size than the original study. Note that disciplines differed in proportion of new versus secondary data replication attempts with business and psychology being mostly new data replications and education and sociology being mostly secondary data replications (see Table S13).

Figure 4. Scatterplot of Pearson's r effect sizes for original and replication outcomes for new data (left) and secondary data (right) replication attempts. New data refers to replication attempts involving data collection (n = 126). Secondary data refers to replication attempts based on existing data (n = 123). Each data point represents the estimated original and replication effect sizes for replicated claims. Size of the bullet is proportional to the number of claims there are per paper to illustrate paper weighting. Replication effect sizes are positive if the observed relationship has the same pattern as the original effect size, and negative if the observed relationship has a different pattern. Data points are classified as successful for effect sizes that achieved statistical significance (p < .05) with the same pattern as the original study and failed for effect sizes that did not.



Discussion

About half of the findings from a sample of social and behavioral science papers published from 2009 to 2018 replicated successfully with variation in success estimates across 13 binary assessments and effect size comparisons. Variation in replicability across the disciplines within the social and behavioral sciences was modest, with replication rates between 42.5% and 49% on the statistical significance metric for fields that had >20 replications. These findings are consistent with the cumulative evidence across systematic replications in the social and

behavioral sciences[3,8,33] and from other fields[12], and they illustrate that there is substantial uncertainty in estimating replicability.

Assessing replicability

There are conceptual, methodological, and inferential challenges to assessing replicability.

Conceptually, it can be challenging to attempt a replication of a prior finding. Strictly speaking, there is no such thing as exact replication. Replications inevitably differ in many ways including the units, treatments, observations, and settings from the original research. Researchers must make decisions about how to conduct a good faith replication of an original claim. For example, should a present-day replication of a 2009 U.S. study of political behavior that used President Obama as a stimulus use Obama again, use the current U.S. president, or use the leader of the participants' nation? The answer depends on what features of that stimulus are essential for testing the original claim. The decision that a new study is a replication of a prior study is a theoretical commitment that they are testing the same claim.[1] The planning and review process of replication studies emphasized making design decisions that would produce a good-faith attempt to replicate the original finding. However, theoretical commitments can be wrong. When ostensible replications produce different results from original studies, it is common and reasonable for the subsequent debate to center on whether it should be considered a replication. The methodology for the replicated studies is available in the supporting information. In particular, deviations identified by the replication team between the replication and the original study and preregistered replication design are extracted from individual reports and highlighted in Tables S4 and S5.

Methodologically, it can be challenging to determine how to measure replication success. We used 13 binary metrics and compared effect sizes. Each approach has features that affect their usability across research designs and statistical models. No singular success metric has been accepted as optimal and universally applicable in the literature.

Inferentially, it can be challenging to determine whether original and replication studies produced the same outcomes. Each of the replication assessment criteria has strengths and weaknesses and may be based on different assumptions. For example, metrics that combine original and replication evidence are not independent tests of replicability and have relatively low sampling error. These tended to suggest the highest success rates. They might be useful only when it is safe to assume no selection or publication bias and the emphasis is on cumulative evidence. Conversely, metrics comparing original and replication effect sizes assume that replication effect sizes significantly smaller and larger than original effect sizes are failures to replicate. These tended to be among the lowest success rates; they might be mostly applicable under conditions in which the precise estimate is important versus knowing the effect is larger than zero. Finally, a reviewer suggested that the metric relying on subjective assessment may be biased by the sample of researchers who participated in this replication project. The raters in this case almost perfectly mimicked using statistical significance for assessing replication success, but other raters might use different criteria. Reasonable minds may disagree on the best way to assess replication success. Productive follow-on investigations will employ this dataset to further evaluate the merits of these metrics.

In sum, the question "did it replicate?" can be difficult to answer. Fortunately, the answer for any given study does not matter much in the long run. Research is conducted on a study by study basis; replicability is established via a cumulative body of evidence. Over time, evidence accumulates and explanations mature. The explanations anticipate and account for variation

across studies and contexts. The importance of deciding whether any two studies showed similar results fades away.

Understanding replicability

Failure to replicate does not mean the original claim was wrong. A single failure to replicate does not justify concluding that the original research was wrong. Even if the replication was perfectly designed, the outcome could be missed or underestimated because of sampling error—a false negative. Even if the replication appeared to be testing the same research question, there could be differences in the methodology, sample, or context that are unrecognized moderators of the outcome. And, even if the replication researchers were diligent in conducting the research, there could be unrecognized errors or flaws in implementing the replication protocol that interfered with observing the outcome.

We attempted to minimize these reasons for failing to replicate by using research designs that were well-powered to detect the original effect size. We also obtained and adapted original materials whenever possible, conducted peer review in advance, preregistered the replication studies, and promoted accountability by committing that materials and data available would be publicly accessible for review to the extent possible. These efforts provide some confidence in the rigor of the replication studies, but do not justify treating the outcomes as sacrosanct.

Successful replication does not mean the original claim was right. A single successful replication does not justify concluding that the original research was correct. The original and replication studies' results could both be observed because of sampling error—a false positive. More importantly, the replicability of an effect is not the same as the validity of its interpretation. Original and replication studies may share confounds, faulty measures, or other design weaknesses that produce replicable, but misinterpreted, outcomes.

The optimal replicability rate is not known. For example, in discovery contexts, it is understood that taking risks on unlikely possibilities will produce many false leads and occasional big rewards. Conducting replications help to reveal weak spots and dead ends, identify boundary conditions, and mature theoretical predictions and explanations that improve replicability over time. In translation of research claims to policy and practice, it may be more important to have established high replicability to have confidence in their applicability and effectiveness.

The problem to solve is not unreplicability per se, it is overconfidence. Published and true are not synonyms[22] and the uncertainty of published claims may be underestimated. For many published findings, it is uncertain whether they will replicate at all, whether they are robust to minor variations in the research context, whether they are generalizable to other contexts, and whether they are valid interpretations of the evidence. Recognition of uncertainty will reduce overconfidence and increase recognition of the value of conducting replications and other verification methods to confront present understanding.[1]

Constraints on generalizability

The sample for this research was a selective representation of the social-behavioral sciences. The inclusion criteria required a positive claim that is supported by a statistical inference. This was practically sensible for the purposes of the program, but it does not cover all relevant research. For example, we excluded research claiming a null result and qualitative research. Our results cannot be expected to generalize to these.

The sample covered a wide range of the social and behavioral sciences. However, the selection of relatively prominent journals might have led to higher or lower replicability than what would be

observed if less prominent journals were also included in the sample. Replicability might have been higher or lower if sampling had reached further back in history, and replicability might be changing in research published after the time frame examined in this project.

Selection effects were minimized within the sample by using stratified random sampling of papers that met the inclusion criteria, but selection effects were introduced in attempting replications because some eligible papers were not matched with a replication team and some replication attempts were not completed (see Methods and SI for more details). The main selection effect was feasibility of conducting a replication given time and cost constraints. It is not difficult to generate plausible hypotheses that original findings from more resource intensive research would be more, less, or similarly replicable as original findings from less resource intensive research.

We did not explore correlations with replication outcomes that might help advance understanding of the reasons for replication success and failure. An initial exploration using this dataset is reported by Abatayo and colleagues in which modest correlations were observed between replication outcomes and several other potential indicators of research credibility.[28] Many other variables could be investigated such as risk of bias in research designs, sample sizes, and evaluation of the impact of differences between original and replication studies.

We also presented outcomes from several different replicability metrics without evaluating their relative merits. A productive line of inquiry would interrogate the relationship between the underlying assumptions of the replicability measures and their impact on observed replicability rates. This will sharpen understanding of what a claim of replication success or failure means, and foster innovation or convergence on how to measure it. The dataset is openly available to stimulate further exploration.

Conclusion

The conditions that promote or inhibit replicability and how to assess it are worthy of additional investigation. Understanding the factors associated with the reliability of evidence will open pathways for advancing theory about research credibility and support pragmatic decision-making for translating research insights into practice.[37]

Methods

This systematic replication effort was part of the SCORE program funded by DARPA to generate and evaluate automated measures of confidence in research claims.[38] Replications provided test data to evaluate the accuracy of human and machine predictions of replicability of claims. Evidence for reproducibility (same analysis, same data) and robustness (different analysis, same data) were also gathered as part of the program. Relations among credibility assessments are reported by Abatayo and colleagues (2025). A full report of the SCORE methodology is accessible through this and several supporting papers.[28,38,39] Data, materials, code, and other outputs from the program are organized and publicly accessible for evaluation and re-use. This methods section summarizes key features of sampling, conducting the replication studies, aggregating the data across replications, and assessment of replication success.

Sampling frame and selection of claims for replication

Claims to replicate were identified with a systematic selection process to reduce selection effects and increase generalizability of the findings to quantitative social and behavioral

research. The project was conducted in two phases. The project started with a sample of 3900 papers selected by a stratified random sampling from a larger set of papers to ensure representativeness across the 62 journals and publication dates from 2009 to 2018. From that pool, 600 papers were randomly selected during Phase 1 as the papers eligible for conducting replication studies with a similar stratified random sampling process to maintain representativeness, and no additional random selection was conducted during Phase 2 to constrain the sample of eligible papers (n = 900). This resulted in a total of 1500 papers eligible for replication with 90.9% of claims subjected to replication attempts selected from the Phase 1 portion. See Abatayo and colleagues [2025] for further details on the sampling frame and selection process.[40]

Eligible papers were matched with research teams with relevant expertise to design and conduct the replication study. Here, random sampling is lost because selection is based on feasibility, available resources, and available expertise. Whenever possible, original methods and materials were collected from the original authors and adapted for the replication study. Replication teams prepared the research design including the methodology and analysis plan and put those through a peer review process that was managed by an independent editor and included independent reviewers plus at least one author of the original study if they agreed to provide review. Authors were instructed to design a good faith replication of the original claim, which could include keeping the methodology the same or updating it in service of improving the quality of the replication attempt. Peer reviewers and editors evaluated the replication design as a holistic assessment of how to improve it to be a good-faith attempt to replicate the original claim. For example, specific instructions for design and evaluation included statements such as “Remember that your goal in replication is to achieve a design that is a good faith test of the original claim (Nosek & Errington, 2020). Sometimes that is a straightforward repeat of the original procedure in your new sample. Sometimes that means adapting the methodology for the new context. Changing the methodology is not bad if it is done so in the service of improving the quality of the replication for testing the original claim.” See the SI for further details on the design and review process.

Replication designs could involve either the collection of new data or finding independent, existing data that was not used for the original research. Approved designs and analysis plans were preregistered on the OSF prior to conducting the research. For the purposes of this project, initiating a draft of the preregistration for the replication study was the milestone defining that the replication had started.

In most cases, a single claim was identified in a single paper and subjected to a single replication attempt with independent data. Of the 1500 papers eligible for replication, 1300 had a single claim isolated for replication and 200 contained additional claims that could be replicated. For 3 claims, multiple replications were conducted using the same protocol, akin to “many labs” studies[4,5,41]. And, for 15 claims, multiple replications were conducted using distinct protocols. For both of these cases, the primary reporting aggregates evidence across multiple replications of a single claim. Finally, there were 27 replications that added new data to data that had been used in the original research. These “hybrid” replications were not included in the main text outcomes because the replications were not independent of the original studies, but they are reported in the SI (Tables S16-S18).

Completed replication reports were reviewed for quality control by team members not involved in the replication study. Data, materials, and code were archived on the OSF and made openly available to the maximum extent allowed without violating privacy of participants or intellectual property licenses for any original materials. A total of 296 replications were conducted and, following aggregation evidence for multiple replications of a single claim there were 274

replications of unique claims from 164 papers. See Figures S1-S11 and Tables S1-S3 and S8-S9 for details about sample selection, study design, attrition, statistical power, and effect size estimation. See Figures S15 and S16 and Table S19 for LLM-generated summaries of the topics and methods represented in the replicated papers.

Replication Assessment Metrics

We assessed the replicability of individual claims from papers that used diverse methodologies. We did this using statistical results of pairs of original and corresponding replication studies. In this section, we describe the approach for each of our 13 binary assessments of replication success.

Statistical Significance and Same Pattern

A common measure of concluding that there is evidence for an original claim and replication of that claim is achieving statistical significance ($p < .05$) with the hypothesized pattern of results. For papers to be included in the sample, the original research needed to have an outcome that could be assessed for replicability with this criterion.

Subjective Interpretation

Subjective assessment of whether the original finding replicated successfully or not, provided by replication teams or project coordinators. No explicit constraints were provided to guide subjective interpretation, and the assessment may be contingent on the identities of the researchers making that subjective judgment. The only reason that this criterion was not used for some findings was because an interpretation was not collected or the interpretation was non-committal to being a success or failure.

Sum of p-values

Calibrating the sum of original and replication *p*-values can control the overall false positive rate and enable replication success even if the original study was non-significant.[42] An unweighted sum of *p*-values concludes that the replication succeeded if the sum of one-sided *p*-values is less than 0.035 (or equivalently, if the sum of two-sided *p*-values is less than 0.07). A weighted version can be used if there are concerns about the diagnosticity of the original evidence, such as the possibility of questionable research practices artificially reducing the *p*-value. We employed the unweighted version, because this method highlights the maximum success rate compared to downweighting the influence of the original study.

Skeptical p-value

This criterion generates a prior using the data from the original result to construct a posterior with an associated credible interval that just overlaps with zero.[43] The skeptical *p*-value assesses the extent to which the replication data is inconsistent with this skeptical prior. The logic is to define in advance how skeptical to be about the replication evidence to believe that the effect does not exist.

Original in the Replication Confidence Interval

This criterion assessed whether the original effect estimate was within the 95% confidence interval of the replication study. This assumes that the original effect was estimated without error and assesses whether it is different from the replication estimate. Replications can produce stronger, weaker, or opposing effects than original studies and fail on this metric.

Replication in the Original Confidence Interval

The complementary criterion is whether the replication effect estimate was within the 95% confidence interval of the original study. This assumes that the replication was estimated without error and assesses whether it is different from the original estimate. Replications can produce stronger, weaker, or opposing effects than original studies and fail on this metric.

Replication in Prediction Interval

The 95% prediction interval has the same basic logic as the approaches using confidence intervals except that it incorporates the precision of both the original and replication effect size in determining the boundaries. As such, this criterion is the most liberal of the interval based methods, including considering some replications estimated near zero or with an opposing pattern to be successful.

Meta-analysis

The fixed-effect meta-analysis criterion combines original and replication evidence into a single estimate and assesses whether the combined evidence is statistically significant with the same pattern as the original study. Because all original studies were positive results, this criterion is necessarily generous to observing replication success as it is not independent of the original evidence.

Bayesian Meta-Analysis

This criterion is the conceptual equivalent of meta-analysis in the Bayesian framework.[44] We used a fixed-effect model to quantify evidence of the effect being present versus absent across both studies. In our implementation the outcome needed to have 'moderate,' 'strong,' or 'extreme' evidence against the null to qualify as a success. The prior for the average effect size is centered at 0 with a standard deviation of 0.25. 1500 iterations per chain were used, with the log-marginal likelihood being estimated by numerical integration with a relative tolerance of 0.1.

Small telescopes

The small telescopes approach assesses whether replication results are consistent with an effect size that could have been detected in the original study.[45] This is calculated in two steps. First, compute the effect size that would have given the original study 33% power. Second, conduct a one-sided hypothesis test of whether the replication data can reject the null hypothesis that it is not smaller than that effect size. This approach recognizes the difficulty of providing evidence for the absence of an effect, so instead defines replication failure as demonstrating that the original study could not have provided evidence for an effect as small as was observed in the replication.

Bayes Factor

The Jeffreys-Zellner-Siow (JZS) Bayes factor is the conceptual equivalent of the standard null hypothesis significance test in the Bayesian framework.[46] It provides relative favorability for the null versus alternative hypothesis, indicating both the absence or presence of an effect. In our implementation the outcome needed to have a Bayes factor against the null of greater than 10 to qualify as a success, corresponding to the interpretation categories of 'strong,' 'very strong,' or 'extreme' evidence.

Replication Bayes Factor

Replication Bayes factor is an alternative to JZS Bayes factor that directly examines the replication evidence in comparison to the original study.[46,47] It provides relative evidence that the replication effect is similar to the original versus being absent. It can only be applied in cases of a non-zero result. In our implementation the outcome needed to have a Bayes factor against the null less than 1 to qualify as a success.

Correspondence test

This criterion considers the correspondence in the effect size estimates between original and replication studies.[48] It combines comparing [a] whether the hypothesis that the effect sizes are the same can be rejected in terms of statistical significance with [b] an equivalence test evaluating whether the hypothesis that the observed difference in effect sizes is not larger than a predefined equivalence threshold. The correspondence test provides four outcomes: [1] *equivalent* = failing to reject that the effect sizes are the same and rejecting that the difference in effect sizes is larger than an equivalence threshold, [2] *trivially different* = rejecting that the effect sizes are the same and rejecting that the difference in effect sizes is larger than an equivalence threshold, [3] *different* = rejecting that the effect sizes are the same and failing to reject that the difference in effect sizes is larger than an equivalence threshold, and [4] *indeterminate* = failing to reject that the effect sizes are the same and failing to reject that the difference in effect sizes is larger than an equivalence threshold. There are enriched possibilities of considering these four outcomes independently using this dataset. For the purposes of creating binary outcomes for comparison with other approaches, we treated *equivalent* and *trivially different* as successful replications, *different* as failed replications, and left out *indeterminate* outcomes.

Data aggregation

For most studies, the original main finding and its corresponding replication findings used the same statistical methods and thus could be assessed on the same effect size scale. However, studies used different statistical methods, and thus also used different effect size scales (e.g., Cohen's *d*, odds ratio, regression coefficient in a multilevel analysis, etc.). These measures cannot be meaningfully compared unless they are converted to a common scale.

We converted as many native effect sizes to partial correlation wherever possible to facilitate these comparisons. Most results could be converted using accepted formulae based on the t, z, or F statistics. In the case of t-statistics, the partial correlation is approximated by t / √(t^2 + residual degrees of freedom). For z statistics, it is approximated by z / √(z^2 + N). And for F statistics, it is approximated by √ ((F * numerator degrees of freedom)/(F* numerator degrees of freedom + denominator degrees of freedom)). Where appropriate, we implemented these using the *effectsize* R package.[49] Some analyses, such as multilevel regression or regressions with clustered standard errors, required a tailored approach to approximate the effective sample size or degrees of freedom for converting to standard effect sizes. In most instances of structural equation models the standardized path coefficients are treated as proxies for the partial correlation, following convention. As needed, partial correlations were also approximated from chi-square statistics or from (log) odds ratios.

The procedures used for each conversion are found in this OSF project (https://osf.io/uqegb/), where the names of each folder correspond to the study ID of the specific replication study. Files with an underscore of "_replication" feature the conversions for the replication findings, while files with an underscore of "_original" feature the conversions for the original findings replicated in that study.

Data analysis and inference

Statistics presented in the paper are largely in the form of descriptive statistics and precision estimates. Proportions of successful replications and similar statistics are aggregated to the paper level unless otherwise noted. Where there are multiple items per paper (e.g., three claims assessed in replication attempts), the sub-level items (e.g., claims assessed nested within papers) are weighted by simple proportion (e.g., each of the three claims receives a weight of ⅓). The code used to generate each statistic reported in this paper is provided in the data and code repository.

All standard errors, confidence intervals, and other metrics of statistical uncertainty are generated by simple clustered bootstrap. Statistical uncertainty for statistics aggregated to the paper level are clustered at the paper level using a clustered bootstrap procedure. 95% confidence intervals are estimated through percentile intervals of the bootstrapped sample distribution.

Inclusion and Ethics

Researchers from 31 nations participated in designing, conducting, and evaluating replications. Joining the collaboration was an open process, promoted via social media primarily by the Center for Open Science and the corresponding author. A variety of roles were defined to maximize opportunity for researchers with varying skills, areas of interest, and access to resources to participate. Criteria for earning co-authorship was defined in advance so that researchers could make informed decisions about joining the collaboration. All replication studies reported in this manuscript involved primary data collection from human participants (Ashland University # 7-22-19-#091, 9-30-19-#105, 9-30-19-#106, 7-2-20#12, 1-31-20#8, 9-8-21#40; Fairfield University # 2712; Western Kentucky University # 20-116, 21-255; University of Nevada Las Vegas # 1521828, 1528491-4; University of San Diego # 2020-70; Occidental College # F19096; University of California, San Diego # 191782SX; University of Exeter # 003030, 001979, 003507, 004097, 004385, 004384, 488230, 488231; London School of Economics and Political Science # 1047; Southern Illinois University # 200071, 21097; University of Toronto # 38581, 38822; Northwestern University # STU00211653, STU00211686; Cornell University # 1912009293, 2001009314, 2105010350, 2105010351, 2109010548; University of Queensland # 2020000052; Texas State University # 7274, 7336; Ruhr University Bochum # 20-6866; University of Texas at Austin # 2020-04-0114; Justus-Liebig-University Giessen # 20-026-757; Ithaca College # 89; Saint Joseph's University # 1522885-1, 1548814-1, 1606422-1, 1606324-1, 1629093-1, 1774195-1; Rochester Institute of Technology # 2112119, 03042120, 02070620, 05041321, 02052721, 01052721; University of Michigan # HUM00173465; University of Pennsylvania # 834860; University of California, Davis # 1547826-1; University of North Carolina, Charlotte # 19-0406, 19-0802, 22-0005; University of Maryland # 1542892-1; Pennsylvania State University # STUDY00013895, STUDY00018137; Vassar College # 01.17.20.01; University of Dayton; Reed College # 2020-S05-FF2, 2020-S30-FF3; University of Texas at Arlington # 2020-0151; University of Groningen # RDMPFEB-20200109-10402; Purdue University # 2020-14; Arizona State University # STUDY00011369; University of Wyoming # 20200113TC02627; Ariel University # AU-SOC-SS-20200122; University of Milan-Bicocca # RM-2020-234; University of New Brunswick # 004-2020, 021-2021, 017-2021; Brigham Young University # 2020-024; University of California, Santa Cruz # 3600; Montclair State University # FY19-20-1652; Carleton University # 112136; Jagiellonian University in Krakow # 1556167-1; Loyola University Chicago # 2908/6640; University of California, Riverside # HS-20-003; Virginia Tech # 20-027; Heinrich Heine University Dusseldorf # 2020-766; University of Cambridge # PRE.2020.011, PRE.2020.086; Rutgers University - Camden # Pro2019002539; University of Chile; Attikon General University Hospital # 136/11-3-2020,

370/7-7-2020, 376/19-7-2021; University of Minnesota # STUDY00009691; University of Connecticut # X20-0102, X21-0162, H21-0079; North Dakota State University # SM20283; Virginia Polytechnic Institute and State University # 20-518; BRANY SBER IRB # 20-041-771, 20-037-770, 20-042-772, 20-032-764, 20-072-839, 20-025-737, 21-066-895) or used secondary analysis of data of organizations, firms, or human participants (University of North Carolina, Charlotte # 19-0804; BRANY SBER IRB # 20-019-749, 21-056-749). All replication studies underwent local ethics review to confirm that the research was performed in accordance with all relevant guidelines and regulations and that informed consent was obtained where necessary. All protocols received concurrence from MRDC HRPO and NIWC-PAC HRPO.

Summary Paragraph

Replication attempts tend to focus on outcomes in a single discipline or topic and with relatively small samples of replication studies. We report a systematic replication of 274 positive claims from 164 papers from 2009 to 2018 from 54 journals to capture diversity across the social and behavioral sciences. Replications were high-powered, used original materials where possible, and were internally peer-reviewed in advance through a standardized protocol. Our work finds statistically significant results with the same pattern as the original study for 151 of 274 claims and for 80.8 of 164 papers weighed for replicating multiple claims per paper. For claims where effect sizes could be converted to Pearson's r, the median effect size saw a 58.1% [95% CI 44.2 - 65.0%] correlation reduction from the original claims to our replication attempts and a 82.4% [95% CI 67.8 - 88.2%] reduction in shared variance. Thirteen methods for evaluating replication success provided estimates ranging from 28.6% to 74.8% with a median of 49.3%. This paper highlights conceptual, methodological, and inferential challenges of determining replicability across the social and behavioral sciences.

Competing Interest Statement

A.H.T., M.D., N.H., K.H., O.M., T.Stankov, B.A.N., and T.M.E. are employees of the non-profit organization Center for Open Science that has a mission to increase openness, integrity, and reproducibility of research.

Data, Materials, and Code Availability Statement

Data, materials, and code associated with this research that can be shared without restriction is publicly available on our OSF repository (https://osf.io/g5sny/). Also included is all available documentation for replication attempts that were not completed. The repository includes a push button package with all code and data used to both produce all statistics, figures, and tables and code that populates them directly into this manuscript from a template. This includes most of the data and code from the individual replication attempts, save for any data that is proprietary or protected that will not be made available, or for which analyst teams were uncertain or unable to confirm that they were allowed to share secondary data. It is possible that some data, materials, or code that could be shared openly is not available at the time of publication. Readers are encouraged to contact the corresponding author or the authors of the relevant subproject (Table S3) to see if more research content can be shared.

Acknowledgements

This work was supported by the Defense Advanced Research Projects Agency (DARPA) under cooperative agreements No: N660011924015 (PI: Brian A. Nosek) and HR00112020015 (PI: Timothy M. Errington). The views, opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. We thank Beatrix Arendt, Alexandria Denis, Mirka Dirzo, Zachary Loomas, Bri Luis, Lesley Markham, E. Simon Parsons, Adam Russell, and Ruben van den Goorbergh for their contributions to this project.

Additional Information

Contact Brian Nosek, nosek@cos.io, for correspondence concerning this paper.

Author ORCIDs and Institutions

Given name	Family Name	ORCID	Institution 1	Institution 2
Andrew H.	Tyner	0000-0001-9180-4490	Center for Open Science	
Anna Lou	Abatayo	0000-0002-2686-5075	Wageningen University and Research	
Mason	Daley	0000-0002-3460-3673	Center for Open Science	
Samuel	Field	0009-0007-4524-2088	SAS Institute	
Nicholas	Fox	0000-0002-3772-8666	Center for Open Science	
Noah A.	Haber	0000-0002-5672-1769	Center for Open Science	
Krystal M.	Hahn	0009-0006-2551-4528	Center for Open Science (COS)	
Melissa	Kline Struhl	0000-0003-2217-9331	Massachusetts Institute of Technology	
Brinna	Mawhinney		Center for Open Science	
Olivia	Miske	0000-0003-4787-3995	Center for Open Science	
Priya	Silverstein	0000-0003-0095-339X	Ashland University	Institute for Globally Distributed Open Research and Education
Courtney K.	Soderberg	0000-0003-1227-7042	Center for Open Science	
Theresa	Stankov		Center for Open Science	
Ahmed	Abbasi	0000-0001-7698-7794	University of Notre Dame	
Christopher L.	Aberson	0000-0003-3481-7177	The Dissertation Coach	
Balazs	Aczel	0000-0001-9364-4988	Institute of Psychology, ELTE Eötvös Loránd University	
Matúš	Adamkovič	0000-0002-9648-9108	Slovak Academy of Sciences	University of Jyväskylä, Finland
Nihan	Albayrak	0000-0003-3412-4311	Boğaziçi University	London School of Economics and Political Science
Peter J	Allen	0000-0002-9690-1545	University of Bristol	
Michael	Andreychik	0000-0002-8542-5504	Fairfield University	
Eli	Awtrey	0000-0002-6712-0256	University of Cincinnati	
Erick	Axxe	0000-0002-0426-5722	Hendrix College	
Flavio	Azevedo	0000-0001-9000-8513	University of Cambridge	University of Utrecht
Miles D.	Bader	0009-0008-2369-2720	Vassar College	
Bence	Bago	0000-0001-6905-1832	Tilburg University	
James	Bailey	0000-0002-6132-6026	Providence College	
Marjan	Bakker	0000-0001-9024-337X	Tilburg University	
Gabriel	Banik	0000-0002-6601-3619	Pavol Jozef Safarik University, Slovakia	
George C.	Banks	0000-0002-2261-2935	University of North Carolina at Charlotte	
Ernest	Baskin	0000-0001-9484-2839	Saint Joseph's University	
Anatolia	Batruch	0000-0002-4669-1463	University of Lausanne	
Annika	Beatteay		University of New Brunswick	
Sophie M.	Behr	0000-0001-5171-671X	DIW Berlin	Technical University of Berlin
Nicholas	Berente	0000-0002-1403-4696	University of Notre Dame	
Zachariah	Berry	0000-0002-0827-6437	University of Southern California	
Jędrzej	Białkowski	0000-0003-1066-7688	University of Canterbury, New Zealand	
Bojana	Bodroža	0000-0003-4165-0678	Faculty of Philosophy, University of Novi Sad	
Laura	Boeschoten	0000-0002-3536-0474	Utrecht University	
Miklos	Bognar	0000-0003-3303-9852	Institute of Psychology, ELTE Eötvös Loránd University	
Christian	Bokhove	0000-0002-4860-8723	University of Southampton	
Diane	Bonfiglio		Ashland University	
Robin	Bouwman	0000-0003-1218-4540	Erasmus University Rotterdam	

Timothy F.	Brady	0000-0001-5924-5211	University of California, San Diego	
Scott	Braithwaite	0000-0002-6180-9765	Brigham Young University	
Gabriel	Briceño Jiménez	0000-0002-3760-9202	Universidad de Chile	
Cameron	Brick	0000-0002-7174-8193	University of Amsterdam	University of Inland Norway
Traci	Bricka	0000-0002-9910-5341	University of Texas at Arlington	
Roman	Briker	0000-0001-6328-7430	Maastricht University	
Annette N.	Brown	0000-0002-4618-8286	FHI 360	
Gordon D A	Brown	0000-0002-2257-1459	University of Warwick	
Robbie	C.M. van Aert	0000-0001-6187-0665	Tilburg University	
Kathryn	Caldwell	0000-0002-2399-7400	Ithaca College	The Analysis Factor
Sara	Capitan	0000-0001-6519-6073	Swedish University of Agricultural Sciences	
Tabaré	Capitán	0000-0002-5055-3995	Swedish University of Agricultural Sciences	
Laura	Caquelin	0000-0003-4557-3315	Karolinska Institutet	
Jesse	Chandler	0000-0001-8151-0915	Mathematica	University of Michigan
Tessa	Charles		Vassar College	
Christopher R.	Chartier	0000-0002-4568-4827	Ashland University	
Rahul	Chawdhary	0000-0001-6260-5935	Kingston University	
Kent Jason	Cheng	0000-0002-8931-4086	University of Massachusetts Boston	
William J.	Chopik	0000-0003-1748-8738	Michigan State University	
Bruce	Clark		Southern Illinois University	
Victoria E.	Colvin		University of Florida	
C. Cozette	Comer	0000-0001-5647-9549	Virginia Tech	
Giulio	Costantini	0000-0001-6610-5452	University of Milan-Bicocca	
Tom	Coupé	0000-0002-9520-5556	University of Canterbury	UCMeta
Jamie	Cummins	0000-0002-9729-4900	University of Bern	
Aneta	Czernatowicz-Kukuczka	0000-0001-7176-8365	Jagiellonian University	
Joshua	de Leeuw	0000-0003-4815-2364	Vassar College	
David	Dobolyi	0000-0002-9493-3447	University of Colorado Boulder	
James N.	Druckman	0000-0002-1249-6790	University of Rochester	
Jianhua	Duan	0000-0002-4750-0243	Stats NZ	University of Canterbury
Marin	Dujmović	0000-0003-1523-227X	University of Bristol	
Daniel J.	Dunleavy	0000-0002-3597-7714	Office of Program Policy Analysis and Government Accountability	
Patrick K.	Durkee	0000-0001-6159-4277	The University of Texas at Austin	California State University, Fresno
Cécile	Emery	0000-0002-7272-1144	University of Exeter	
Kevin M.	Esterling	0000-0002-5529-6422	University of California Riverside	
Thomas R.	Evans	0000-0002-6670-0718	University of Greenwich, UK	
Anna	Fedor	0000-0003-2290-2323	independent researcher	
Belén	Fernández-Castilla	0000-0002-3451-0637	Universidad Nacional de Educación a Distancia	
Nathan	Fiala		University of Connecticut	RWI - Leibniz Institute for Economic Research
James G.	Field	0000-0001-8487-6648	West Virginia University	
Nathan	Fong	0000-0001-7891-3262	Rutgers University	
Miguel A.	Fonseca	0000-0002-5294-6784	University of Exeter	
Alexandra L.J.	Freeman	0000-0002-4115-161X	University of Cambridge	
Jeremy	Freese		Stanford University	
Sandra J.	Geiger	0000-0002-3262-5609	University of Amsterdam	University of Vienna

Jing	Geng	0000-0002-7059-7725	Virginia Tech	
Laura M.	Getz	0000-0002-3429-7506	University of San Diego	
Linda Marjoleine	Geven	0000-0001-5075-5223	Leiden University	
Ilka Helene	Gleibs	0000-0002-9913-250X	London School of Economics	
Donna Pamella	Gonzales	0009-0006-7076-661X	NA	
Janaki	Gooty	0000-0002-9447-0454	UNC, Charlotte	
Amélie	Gourdon-Kanhukamwe	0000-0002-3060-1320	Kingston University	King's College London
Cristina	Greculescu	0000-0003-0384-552X	Bremen International Graduate School of Social Sciences (BIGSSS)	Constructor University
Siobhán M.	Griffin	0000-0002-3613-2844	University of Limerick, Ireland	
Lusine	Grigoryan	0000-0002-2077-1975	University of York	Ruhr University Bochum
Martina	Grunow	0000-0003-2617-9638	Leuphana University of Lüneburg	
Nicholas	Gunby	0009-0001-6003-9068	Contact Energy	UCMeta
Braeden	Hall	0000-0002-4157-3054	Southern Illinois University Carbondale	
Paul H. P.	Hanel	0000-0002-3225-1395	University of Essex	University of Bath
Erin E.	Hannon	0000-0002-3482-5954	University of Nevada Las Vegas	
Sam	Harper	0000-0002-2767-1053	McGill University	
Marco Jürgen	Held	0000-0003-2370-2905	University of Bamberg	
Louis	Hickman	0000-0002-2752-7705	Virginia Tech	
Nathan C.	Higgins	0000-0002-1219-6448	University of Nevada Las Vegas	University of South Florida (current)
Svenja	Hippel	0000-0002-5447-3730	University of Bonn	
Sven	Hoeppner	0000-0003-2697-4420	Charles University	Freie Universität Berlin
Sanghyun	Hong	0000-0003-0135-2617	University of Canterbury	
Thomas J.	Hostler	0000-0002-4658-692X	Manchester Metropolitan University	
Michael	Inzlicht	0000-0001-9297-6497	University of Toronto	Rotman School of Management
Kamil	Izydorczak	0000-0002-9870-3825	SWPS University	
Bastian	Jaeger	0000-0002-4398-9731	Tilburg University	
Kristin	Jankowsky	0000-0002-4847-0760	University of Kassel	
Johannes	Jarke-Neuert	0000-0001-6407-5202	Forschungszentrum Jülich	Universität Hamburg
Matthew	Jensen	0000-0001-8711-1827	University of Oklahoma	
Biljana	Jokić	0000-0002-0829-4037	University of Belgrade, Faculty of Philosophy	Metropolitan University, Belgrade
Daniel	Jolles	0000-0003-1277-0793	London School of Economics and Political Science	University of Essex
Phillip	Jolly	0000-0001-7835-3613	Pennsylvania State University	
Angela M.	Jones	0000-0001-7605-4206	Texas State University	
Marie	Juanchich	0000-0003-4256-4501	University of Essex	
Pavol	Kačmár	0000-0003-0076-1945	Faculty of Arts, Pavol Jozef Šafárik University in Košice	
Hansika	Kapoor	0000-0002-0805-7752	Monk Prayogshala	University Of Connecticut
Andjela	Keljanovic	0000-0002-0562-394X	University of Pristina-Kosovska Mitrovica, Faculty of Philosophy	University of Novi Sad, Faculty of Philosophy
Samjhana	Koirala		University of Connecticut	
Marta	Kołczyńska	0000-0003-4981-0437	Institute of Political Studies of the Polish Academy of Sciences	
Dimitra	Kouroupaki		Atticon University General Hospital, 2nd Psychiatric Clinic	
Ulrich	Kühnen	0000-0001-9059-4719	Constructor University, Bremen, Germany	

Michelangelo	Landgrave	0000-0003-4348-380X	University of Colorado, Boulder	University of California, Riverside
Michael J.	Larson	0000-0002-8199-8065	Brigham Young University	
Lyonel	Laulié	0000-0001-9817-4110	Universidad de Chile	
Alice C E	Lawrence	0000-0002-8771-4844	University of Cambridge	
Joel M.	Le Forestier	0000-0003-3330-2171	University of Pittsburgh	
Katelin E.	Leahy	0000-0002-3638-3694	Michigan State University	
Sungmok	Lee		University of New Brunswick	
Jared	Leslie		Univeristy of Nevada Las Vegas	
Savannah C.	Lewis	0000-0002-9948-1195	Ashland University	University of Alabama
Christopher	Limnios	0000-0001-5387-1334	Providence College	
Hause	Lin	0000-0003-4590-7039	Cornell University	Massachusetts Institute of Technology
An-Chiao	Liu	0000-0003-4064-0515	Utrecht University	
John Wills	Lloyd	0000-0002-2597-6216	University of Virginia	
Elliot A	Ludvig	0000-0002-0031-6713	University of Warwick	
Dermot	Lynott	0000-0001-7338-0567	Maynooth University	
Jordan	MacDonald	0000-0001-6687-3638	University of New Brunswick	
Peter	Mallik	0000-0003-3429-2198	Hubbard Decision Research	
Daniel J.	Mallinson	0000-0002-8094-6685	Penn State Harrisburg	
Daniele	Marinazzo	0000-0002-9803-0122	Ghent University	
Corinna S.	Martarelli	0000-0001-9160-793X	UniDistance Suisse	
Joshua	Matacotta	0000-0002-0192-5254	Western University of Health Sciences	Integrated Behavioral Health Research Institute
Andrew	McBride	0000-0002-0810-9751	Santa Clara University	
Cillian	McHugh	0000-0002-9701-3232	University of Limerick	
Gail	McMillan	0000-0001-5619-5618	Carleton University	
Esteban	Méndez	0000-0002-7248-6092	Central Bank of Costa Rica	
Mitchell	Metzger		Ashland University	
Michalis P.	Michaelides	0000-0001-6314-3680	University of Cyprus	
Johannes	Michalak	0000-0003-4701-5464	Witten/Herdecke University	
Leticia	Micheli	0000-0003-0066-8222	Julius-Maximilian University of Würzburg	Leiden University
Jeremy K.	Miller	0000-0003-4409-7660	Willamette University	
Marina	Milyavskaya	0000-0002-0510-4891	Carleton University	
Daniel C.	Molden	0000-0002-2182-5621	Northwestern University	
Ambar G.	Monjaras		University of Nevada, Las Vegas	University of Nevada, Las Vegas
David	Moreau	0000-0002-1957-1941	University of Auckland	
Audrey	Morrow	0000-0001-9143-0708	University of California Santa Cruz	
Cristóbal	Moya	0000-0002-7176-4775	DIW Berlin	Universität Bielefeld
Liad	Mudrik	0000-0003-3564-6445	Tel Aviv University	
Laetitia B.	Mulder	0000-0003-3203-7601	University of Groningen	
Katie A.	Munt	0000-0002-3823-2005	The University of Queensland	
Arijit	Nandi	0000-0002-3399-0536	McGill University	
Kathryn	Nason	0000-0001-6882-5517	University of New Brunswick	
Carolin	Nast	0000-0003-1067-2946	UiS Business School	
Gideon	Nave	0000-0001-6251-5630	The Wharton School of Business, University of Pennsylvania	
Heinrich H.	Nax	0000-0003-1261-8134	University of Zurich	ETH Zurich
Florian	Neubauer	0000-0001-7134-3439	University of Connecticut	RWI - Leibniz Institute for Economic Research
Phuong Linh L.	Nguyen	0000-0002-1575-0395	University of Minnesota	

Austin Lee	Nichols	0000-0003-4580-3301	Central European University	
Gustav	Nilsonne	0000-0001-5273-0150	Karolinska Institutet	Stockholm University
Ernest	O'Boyle	0000-0002-9365-1069	Indiana University	
Jule	Oettinghaus		Ruhr-Universität Bochum	
Jeewon	Oh	0000-0001-8103-906X	Syracuse University	
Adoril	Oshana	0000-0003-2386-826X	University of North Carolina at Charlotte	
Thomas	Ostermann	0000-0003-2695-0701	Witten/Herdecke University	
Rachel P.	Ostrowski	0009-0006-3953-6955	Vassar College	
Abiola	Oyebanjo		Policy Innovation Center	
Radoslaw	Panczak	0000-0001-5141-683X	University of Bern	
Jamie	Patrianakos	0000-0002-0198-844X	Loyola University Chicago	
Ignacio	Pavez	0000-0001-5257-5330	Universidad de Chile	
Yuri G.	Pavlov	0000-0002-3896-5145	University of Tuebingen	
Sofia	Persson	0000-0002-7353-5204	Leeds Beckett University	
Marco	Perugini	0000-0002-4864-6623	University of Milan Bicocca	
Kim	Peters		University of Exeter	
Constant	Pieters	0000-0001-5567-6401	Copenhagen Business School	
Vladimir	Ponizovskiy	0000-0002-1592-5482	Ruhr-Universität Bochum	Durham University
Nathaniel D.	Porter	0000-0002-0479-6777	Virginia Tech	
Jason M.	Prenoveau	0000-0001-9388-7410	Loyola University Maryland	
Danka	Purić	0000-0001-5126-3781	University of Belgrade, Faculty of Philosophy	
Mariah F.	Purol	0000-0003-2921-3600	Union College	
Arathy	Puthillam	0000-0003-2426-8362	Monk Prayogshala	UC San Diego
Kimberly A.	Quinn	0000-0002-0751-0172	DePaul University	
Marco	Ramljak	0009-0008-1502-6453	Utrecht University	
W. Robert	Reed	0000-0002-6459-8174	University of Canterbury	UCMeta
Michaela	Ritchie	0000-0001-9872-3220	University of New Brunswick Saint John	
Margaret	Ritzau		Vassar College	
Sean Patrick	Roche	0000-0003-2662-2716	Texas State University	
Romina	Rodela	0000-0003-0070-6794	Södertörn University	
Jan Philipp	Röer	0000-0001-7774-3433	Witten/Herdecke University	
Ivan	Ropovik	0000-0001-5222-1233	Charles University	Czech Academy of Sciences
Jacob	Rothschild	0000-0002-8416-9879	Verasight	
Justine	Saal		Ruhr-Universität Bochum	
Hani	Safadi	0000-0003-0609-8005	The University of Georgia	
Jason	Samaha	0000-0001-8010-5993	University of California, Santa Cruz	
Mary	Sanchez	0009-0009-7217-7538	University of Nevada, Las Vegas	
Soorya	Sankaran	0000-0001-6169-3080	University of California, Santa Cruz	
David	Santos	0000-0001-9786-5219	Universidad Autónoma de Madrid	
Amanda C.	Sargent	0000-0003-4599-622X	Bentley University	
Marian	Sauter	0000-0003-3123-8073	Universität Ulm	
Kathleen	Schmidt	0000-0002-9946-5953	Southern Illinois University	Ashland University
Landon	Schnabel	0000-0002-2674-3019	Cornell University	
Amber N	Schroeder	0000-0002-6955-4322	The University of Texas at Arlington	
Sebastian W.	Schuetz		University of Colorado Boulder	
Brendan A.	Schuetze	0000-0002-5210-6785	The University of Texas at Austin	University of Potsdam

Michael	Schulte-Mecklenbeck	0000-0002-0406-8809	University of Bern	Max Planck Institute for Human Development
Astrid	Schütz	0000-0002-6358-167X	University of Bamberg	
Eric L.	Sevigny	0000-0002-1596-0042	Georgia State University	
Ellie	Shackleton	0009-0008-4044-3558	University of Limerick	
Richard M.	Shafranek	0000-0002-8055-4008	Northwestern University	
Samuel	Shaki	0000-0002-2340-5401	Ariel University	
Shishir	Shakya	0000-0002-6272-6654	Appalachian State University	
Miroslav	Sirota	0000-0003-2117-9532	University of Essex	
Matthew Ryan	Sisco	0000-0003-0296-7664	Columbia University	
Maksim M.	Sitnikov	0009-0005-0904-7395	Tilburg University	
L. Robert	Slevc	0000-0002-5183-6786	University of Maryland, College Park	
Laura	Smalarz	0000-0002-2435-7843	Arizona State University	
Colin Tucker	Smith	0000-0002-8132-9988	University of Florida	
Joel S.	Snyder	0000-0002-5565-3063	University of Nevada, Las Vegas	
Nicolas	Sommet	0000-0001-8585-1274	University of Lausanne	
Fatih	Sonmez	0000-0002-4054-0269	Muş Alparslan University	
Barbara A.	Spellman	0000-0003-2823-1292	University of Virginia	
Natalia	Stanulewicz-Buckley	0000-0003-3672-3422	Aston University	
George	Stock	0000-0002-6391-2731	UNC Charlotte	
Chris N. H.	Street	0000-0002-0416-485X	Keele University	
Eirik	Strømland		Western Norway University of Applied Sciences	
Tina	Sundelin	0000-0002-7590-0826	Stockholm University	Karolinska Institutet
Moin	Syed	0000-0003-4759-3555	University of Minnesota	
Anna	Szabelska	0000-0001-5362-3787	Psychological Science Accelerator	
Barnabas	Szaszi	0000-0001-7078-2712	Institute of Psychology, ELTE Eötvös Loránd University	
Ewa	Szumowska	0000-0002-4181-6175	Jagiellonian University	University of Maryland, College Park
Anirudh	Tagat	0000-0002-7707-453X	Monk Prayogshala	
Susanne	Täuber	0000-0003-2859-1474	University of Amsterdam	
Louis	Tay	0000-0002-5522-4728	Purdue University	
Stuti	Thapa	0000-0002-0740-7204	University of Tulsa	
Jason	Thatcher	0000-0002-7136-8836	Temple University	Colorado University-Boulder
Domna	Tsaklakidou		ATTIKON' University General Hospital	
Lars	Tummers	0000-0001-9940-9874	Utrecht University	
Elise	Turkovich	0000-0002-7333-9190	University of California, Santa Cruz	
Melba Verra	Tutor	0000-0001-7951-3690	Independent researcher	
Karolina	Urbanska	0000-0001-5063-4747	Independent Researcher	
Anna Elisabeth	van 't Veer	0000-0002-2733-1841	Institute of Psychology, Leiden University	
Marcel	van Assen	0000-0002-7517-6081	Tilburg University	Utrecht University
Niels	van de Ven	0000-0002-6730-9200	Tilburg University	
Ruben	van den Goorbergh	0000-0003-3229-3015	Utrecht University	
Elisabeth Julie	Vargo	0000-0002-5123-1170	Institute for Globally Distributed Open Research and Education (IGDORE)	
Leigh Ann	Vaughn	0000-0002-2399-7400	Ithaca College	
Simine	Vazire	0000-0002-3933-9752	University of Melbourne	

Jentien M.	Vermeulen	0000-0002-6245-6498	Amsterdam UMC location AMC	
Diem Thi Hong	Vo	0000-0002-5289-2325	RMIT University Vietnam	UCMeta
Victor	Volkman	0000-0003-2781-535X	University of Connecticut	
Eric-Jan	Wagenmakers	0000-0003-1596-1034	University of Amsterdam	
Deliah	Wagner	0000-0002-1508-3190	Center for Criminological Research Saxony (ZKFS)	University of Jena
Lukasz	Walasek	0000-0002-7360-0037	University of Warwick	
Frank	Walter		Justus-Liebig-University Giessen	
Lara	Warmelink	0000-0003-1218-9448	Lancaster University	
Liuqing	Wei	0000-0001-6488-7454	Hubei University	
Marie Isabelle	Weißflog	0000-0003-0686-0581	Ruhr University Bochum	
Nicholas	Weller	0000-0002-2198-5625	University of California, Riverside	
Aaron L.	Wichman	0000-0002-2641-440X	Western Kentucky University	
Jonathan	Wilbiks	0000-0002-6882-5215	University of New Brunswick	
Jamal R.	Williams	0000-0002-3034-511X	University of California, San Diego	
Kelly	Wolfe	0000-0002-4077-6415	Heriot-Watt University	
Finnian	Wort	0000-0002-7618-7700	University of Warwick	
Ryan	Wright	0000-0002-9719-415X	University of Virginia	
Jesper N.	Wulff	0000-0002-7976-0939	Aarhus University	
Xindong	Xue	0009-0000-6959-0924	Zhongnan University of Economics & Law	
Veronica X.	Yan	0000-0002-3988-3184	The University of Texas at Austin	
Yuzhi	Yang	0009-0005-2091-6299	University of New Brunswick	
Sangsuk	Yoon	0000-0002-3399-1096	University of Dayton	
Iris	Žeželj	0000-0002-9527-1406	University of Belgrade, Serbia	
Yinxian	Zhang	0000-0002-5343-5898	Queens College, CUNY	
Ignazio	Ziano		University of Geneva	
Cristina	Zogmaister	0000-0002-1540-7503	Università di Milano-Bicocca	
Zorana	Zupan	0000-0002-0763-8192	University of Belgrade	
Rolf A.	Zwaan	0000-0001-9967-7879	Erasmus University Rotterdam	
Brian A.	Nosek	0000-0001-6797-5476	Center for Open Science	University of Virginia
Timothy M.	Errington	0000-0002-4959-5143	Center for Open Science	

Author Contributions: CReDiT Taxonomy

Full name	Conceptualization	Data curation	Formal Analysis	Funding acquisition	Investigation	Methodology	Project admin	Resources	Software	Supervision	Validation	Visualization	Original draft	Review & editing	Contribution Explanation
Andrew H. Tyner	1	1	0	0	1	1	1	0	0	1	1	1	1	1	I was a Research Scientist and then Principal Research Scientist for the duration of SCORE. In these roles I was involved in most aspects of implementing the project, including claim extraction, facilitating replications and reproductions, quality assurance, data curation, analysis, visualizations, and oversight responsibilities.
Anna Lou Abatayo	0	0	0	0	1	1	0	0	1	0	1	0	0	0	Extracted claims, mostly from economics journals but helped out on non-economics journals / covid-preprints when needed. Reviewed others' extracted claims. Part of the team (along with Andrew Tyner) that created the process to replicate extracted claims. Collected data for replication (also collected data for reproduction when replication and reproduction data overlapped). Analyzed data for replication. Managed excel file where external individuals picked a claim / journal article to replicate. Read through and checked pre-analysis plans for replication by others before they were submitted. Read through and checked the analysis (for replication). Recruited external individuals that helped with replication (and they might have helped for reproduction too).
Mason Daley	0	0	0	0	1	0	0	0	0	0	0	0	0	1	PR coding
Samuel Field	1	0	0	0	1	1	0	0	0	0	0	0	0	0	I collaborated in the development and execution of research methodology related to the identification of research claims in sampled articles. I also conducted preliminary power analyses for the sub-sample of studies selected for replication.
Nicholas Fox	0	0	0	0	1	1	0	0	0	0	0	0	0	0	Research scientist in TA1, coding scientific papers to generate datasets for TA2 and TA3 usage, as well as managing research laboratories conducting replication attempts for validation
Noah A. Haber	0	1	1	0	0	1	0	0	0	0	1	1	0	1	Assisted in the analysis portion of this project, contributing primarily to analysis code, data visualizations, data management, manuscript reproducibility, and methods.
Krystal M. Hahn	0	0	0	0	1	0	0	0	0	1	0	0	0	0	I coordinated IRB and HRPO review, submissions, and approvals. I also assisted with preregistration templates and review. I contributed to non-HSR coordination and sourcing, as well as the OSF audit.
Melissa Kline Struhl	1	0	1	0	1	1	0	0	1	1	0	0	0	0	I was one of four research scientists on the COS SCORE team, involved in designing and executing processes for SCORE claim selection, replication/reproduction design, data management & analysis of primary outcomes of finished replications/reproductions
Brinna Mawhinney	0	0	0	0	1	1	0	0	0	0	0	1	1	1	As a Project Coordinator with Center for Open Science, I have helped with the development of coding schemes relevant to SCORE's methodology, conducted data collection, created charts/graphs for presentations about SCORE, and will be contributing to writing and review of publications.
Olivia Miske	0	0	0	0	1	1	1	0	0	1	1	0	1	1	I was a Project Coordinator (and later Assistant Research Scientist) for SCORE at COS. I was involved in a range of activities which included: communicating with replication labs and managing coordination/tracking throughout the process; assisting with the preregistration review process; managing the IRB/HRPO submission and review process; conducting power analyses for a subset of replication studies; data entry/validation of replication outcomes for a subset of projects; assisting with the OSF audit process, contributing to the methods write-up.
Priya Silverstein	0	0	0	0	0	1	0	0	0	0	1	0	0	1	I was involved in lots of different bits and bobs! CE (single-trace and bushel), had input on some P2 process stuff (especially bushel CE), variable coding (original, replication, reproduction), and some validation stuff (OSF audits, checking CE and variable coding, etc.)
Courtney K. Soderberg	0	0	0	0	1	1	0	0	0	1	0	0	0	0	While at COS, developed practices and methodologies for project's original paper statisical coding and power and sample size analysis and planning. I implemented them myself as well as trained and supervised other research scientists and associates on the application of the processes and analyses.

Theresa Stankov	0	0	0	0	0	0	0	0	1	0	0	0	0	0	Set up raw data to processed data pipeline for replication outcomes
Ahmed Abbasi	0	0	0	1	1	0	0	0	0	0	0	0	0	0	I was part of the investigation team for "Replication of a Research Claim from Johnson et al. (2012), from The Journal of Organizational Behavior." I worked closely with David Dobolyi at CU-Boulder (he was involved with investigation and also performed formal analysis), Ryan Wright at UVa (McIntire School of Commerce), and others. My role was data collection (at Notre Dame and UVa) for Stage 1 and 2. I also worked with collaborators at UVa to secure funding for our replication (to compensate student subjects) through the McIntire Center for Business Analytics at UVa. See final report for details.
Christopher L. Aberson	0	0	1	0	0	0	0	0	0	0	0	0	0	0	Statistical power and sample size requirement calculations
Balazs Aczel	0	0	0	0	1	0	0	0	0	0	0	0	0	0	We conducted a replication research for Mason et al. 2012
Matúš Adamkovič	0	0	1	0	1	1	0	0	0	0	0	0	0	1	Together with my colleagues Ivan Ropovik and Gabriel Baník, we carried out a replication of Stahl_JournPerSocPsy_2012_gbl9 (393) in SCORE Phase 1. In Phase 2, we conducted a data analytic replication of BERSANI_Criminology_2013_zmYY_6zz3k.
Nihan Albayrak	0	0	1	0	1	1	0	0	1	0	1	0	0	1	I planned and conducted a replication.
Peter J Allen	0	0	1	0	1	1	0	0	0	0	0	0	0	0	Ran one of the replication studies.
Michael Andreychik	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I completed a replication of Nyhan and Reifler (2015). This included creating the replication materials and programming the software to conduct the replication, analyzing the results of the replication, and writing a report of the methods and results. I worked closely with Andrew Tyner and Zack Loomis of COS on the project.
Eli Awtrey	0	0	0	0	0	0	0	0	0	0	1	0	0	0	Reviewer for reproduction of claims from (1) Smith et al. (2016), (2) Fox (2009), (3) Denson & Chang (2009), (4) Fitzgerald et al. (2018), (5) Wise et al. (2020), (6) Teovanovic et al. (2020)
Erick Axxe	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I conducted one replication study. (O'Brien_AmSocioRev_2015_7X54 - Ramljak/Axxe - Data Analytic Replication - 93k7)
Flavio Azevedo	1	0	1	1	1	1	0	0	1	1	1	1	1	1	I was the lead/PI for the Replication of a research Claim from Ihme & Tausendpfund (2018) from The Journal of Experimental Political Science Ihme_JournExpPoliSci_2018_xYbO_yk16. The other two members were Leticia Micheli and Deliah Bolesta.
Miles D. Bader	0	0	0	0	1	1	0	0	1	0	0	0	0	0	I developed methodology, designed software, and executed data collection for a replication study and a generalization study, Rich & Gureckis and Griffiths & Tenenbaum respectively, in collaboration with Rachel Ostrowski and under the supervision and tremendous guidance of Josh de Leeuw in his Cognitive Science lab at Vassar College.
Bence Bago	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I completed 1 replication research; formal analysis, software, methodology, and investigation as well.
James Bailey	0	0	1	0	1	1	0	0	1	0	1	0	0	0	Prepared data for 2 replications, and served as a reviewer for 2 replications.
Marjan Bakker	0	0	1	0	0	0	0	0	0	0	0	0	0	1	I was involved in the power analysis group
Gabriel Banik	0	0	0	0	1	0	0	0	0	0	0	0	0	1	Together with my colleagues Matus Adamkovic and Ivan Ropovik, we carried out a replication of Stahl_JournPerSocPsy_2012_gbl9 (393) in SCORE Phase 1. In Phase 2, we conducted a data analytic replication of BERSANI_Criminology_2013_zmYY_6zz3k.
George C. Banks	0	0	1	0	1	1	0	0	0	1	1	0	0	1	I led and participated in the completion of two replication studies. The first was a primary data collection using a survey-based design. The second was a meta-analysis replication.
Ernest Baskin	0	0	1	0	1	1	0	0	1	1	1	0	0	0	I need a number of replications across 2 waves of the project. My recollection is that there were at least 15 replications. Replications were completed and shepherded through peer review by myself alone. I also participated in areas where I replicated the analyses of original papers.
Anatolia Batruch	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I conducted the replication of a sociological study with Nicolas Sommet
Annika Beatteay	0	0	1	0	1	1	0	0	1	0	0	0	0	0	
Sophie M. Behr	0	0	0	0	1	1	0	0	1	0	0	0	0	0	I did data preparation work for 2 replication studies: Lersch_EurSocioRev_2014_LJXK and Bro_ckel_EurSocioRev_2015_PyYd.
Nicholas Berente	0	0	0	0	1	0	0	0	0	0	0	0	0	0	Helped with data collection at the University of Notre Dame and the University of Georgia (my previous institution)

Zachariah Berry	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I conducted 4 replication studies to replicate a variety of effects within 4 different published articles.
Jędrzej Białkowski	0	0	0	0	1	0	0	0	0	0	1	0	0	0	Planned and conducted a replication 1 & Prepared data for replication 2
Bojana Bodroža	0	0	0	0	1	1	0	0	0	0	0	0	0	1	I participated in the study Luttrell (2016). I participated in writing the preregistration of the study. I gathered data (student sample from my university) i.e. a part of the sample. I also wrote the SCORE report which is uploaded to the OSF.io project page.
Laura Boeschoten	0	0	0	0	0	0	0	0	0	0	1	0	0	0	I was a data finder as part of the replication of a Research Claim from Böhnke & Link (2017), from European Sociological Review (see https://osf.io/w56ve/)
Miklos Bognar	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I conducted a replication for Mason et al. (2012) with all of its tasks, like programming the experiment software, collecting and analysing data and reporting the results.
Christian Bokhove	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I performed a (conceptual) replication of the article by Kim et al. (2014) with secondary data from the IEA's 2018 International Computer and Information Literacy Study. I created the statistical models, created the R-code, and provided interpretation of the findings. Apart from the OSF repository, the results have also been published in https://www.tandfonline.com/doi/full/10.1080/13803611.2021.2022319
Diane Bonfiglio	0	0	0	0	1	0	0	0	0	0	0	0	0	0	Along with my colleague Peter Mallik, I collected data to perform a replication of Bursztyn (2017).
Robin Bouwman	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I designed and programmed 1 replication study for lab testing. Given COVID lockdowns, i re-programmed the experiment for online fielding. I ran the data collection, and the data analysis (R-code) and presented the results in the results report. I collaborated with Prof. dr. Lars Tummers from Utrecht university, the Netherlands.
Timothy F. Brady	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I conducted a replication study of the research claim from 'Asymmetrical Body Perception: A Possible Role for Neural Body Representations', by Linkenauger et al. (2009).
Scott Braithwaite	0	0	0	0	1	0	0	0	0	0	1	0	0	0	My colleagues and I conducted one of the replication submissions. I served as a reviewer for multiple replications submissions.
Gabriel Briceño Jiménez	0	0	1	1	1	1	0	0	0	0	1	1	1	1	I assisted in acquiring funding for the project by preparing all the necessary materials and documents to be considered for the SCORE project. I was responsible for obtaining the Ethical Protocol approvals from both the ethics committee in my country and the United States. Additionally, I developed various data collection instruments and coordinated data collection with participants. Lastly, I handled tasks such as cleaning databases, conducting data analysis, and drafting the 'Data Collection and Procedures' and 'Results and Analysis' sections. This work can be checked in the following paper: https://iamr.uchile.cl/index.php/EDA/article/view/66826
Cameron Brick	0	0	0	0	0	0	0	0	0	0	1	0	0	1	I reviewed two pre-registration designs.
Traci Bricka	0	0	1	0	1	1	0	0	0	0	0	0	0	0	Served as a co-researcher alongside Amber Schroeder on the Nakai et al. (2011) replication project.
Roman Briker	0	0	1	0	1	1	0	0	1	0	0	0	0	1	I (together with Frank Walter) replicated a study from applied psychology (Reh et al., 2018). I was involved in planning the study, collecting the data, analyzing the data, and writing up and reporting the data back to COS/SCORE.
Annette N. Brown	0	0	0	0	0	0	0	0	0	1	1	0	0	0	Served as a SCORE replications editor and supervised the reviews for 18 replication protocols including compiling and reconciling external reviews, formulating questions and requirements for replication researchers, making final determinations on replication protocols.
Gordon D A Brown	0	0	1	0	1	1	0	0	0	0	0	0	0	0	Member (with Lukasz Walasek, Elliot A. Ludvig, Finnian Wort) on: Replication of a research claim from Bart de Langhe, Stijn, M. J. van Osselaer, Stefano Puntoni and Ann L. McGill (2014) in Journal of Consumer Research.
Robbie C.M. van Aert	0	0	1	0	0	0	0	0	1	1	0	0	0	1	I was part of the team from Tilburg University that conducted statistical power analyses and computed variables (e.g., statistical power, effect size, etc) for the original and replication studies. I also created R code for computing these variables. I expect to do some reviewing and editing for the paper.
Kathryn Caldwell	0	0	1	0	1	1	0	0	1	0	0	0	0	0	Leigh Ann Vaughn and I developed the replication of the research claim from Ku and Zaroff (2014), from the Journal of Environmental Psychology. Our Ithaca College subcontract number was ithaca_4zy8. We designed the replication study, developed and revised the preregistration, including the initial Department of the Navy (DON) Human Research Protection Office (HRPO), collected and analyzed data, and submitted and revised our final report for SCORE.
Sara Capitan	0	0	0	0	1	1	0	0	0	0	0	0	0	0	I worked on a replication.

Laura Caquelin	0	0	1	0	0	0	0	0	0	0	1	0	0	0	Provided review of analysis code for ensuring clarity, accuracy, and reproducibility.
Tabaré Capitán	0	0	1	0	1	1	0	0	1	1	1	0	0	0	I might be wrong, this is from a quick look at my folder. Data analyst (5x), data finder (1x), reviewer (5x), formal replication (1x), and editor (10x).
Jesse Chandler	0	0	1	0	1	1	0	0	1	0	0	0	0	1	I conducted the replication of Bartels and Urminsky (2015).
Tessa Charles	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I worked with Joshua de Leeuw on replications for Griffiths & Tenenbaum (2011) and van Dijck, Gevers, and Fias (2009), as both a researcher and author.
Christopher R. Chartier	0	0	1	0	1	1	0	0	1	1	0	0	0	1	Recruiting and matching teams to conduct replication studies. Also part of two-person team (with Savannah C. Lewis) to conduct one of the replication studies.
Rahul Chawdhary	0	0	0	0	0	0	0	0	0	0	1	0	0	0	I validated 2 projects.
Kent Jason Cheng	0	0	1	0	1	0	0	0	1	0	0	0	0	0	I was involved in data collection for four different replication assignments (Fielding-Miller study, Montez et al study, Fitzgerald et al study, and Carrillo Vega study). I worked with Daniel Mallinson on publishing the replication results for the Fitzgerald study.
William J. Chopik	1	0	1	0	1	1	0	0	0	1	1	0	1	1	For our replication project (Nelson), I (with help from students) programmed the survey and edited/spliced/hosted the videos for the study. I coordinated data collection and training of RAs, and supervised data analysis/reporting (which was mostly done by the students). I reviewed several replication/pre-registrations. I encouraged those students (Mariah Purol, Jeewon Oh, Katelin Leahy) to complete this form.
Bruce Clark	0	0	0	0	0	0	0	0	0	0	0	0	1	1	Edited and reviewed at least three drafts.
Victoria E. Colvin	0	0	1	0	1	0	0	0	1	0	0	0	0	0	I conducted the replication of one project along with. Dr. Colin Smith. We replicated that claim from Study 1 of Axt et al., (2018) that members of a nondominant racial group will show reliable implicit ingroup preference when presented with positively valanced information, and will show reliable implicit preference for the dominant group when presented with negative valanced information. My role in conducting this replication was to help write code on Project Implicit, analyze the data, write up the results, and upload and organize all documents on the OSF page.
C. Cozette Comer	0	0	0	0	1	0	0	0	0	0	0	0	1	1	Nathaniel D. Porter and I conducted a replication of Raver & Nishii (2010) I served as a pre-registration peer reviewer for two projects: Kuo & Raley (2016) Demography; Humphrey (2017) SocSciMed
Giulio Costantini	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I conducted and analyzed 1 replication, Christensen et al. (2018) published in the European Journal of Personality, project code Christensen_EurJournPersonality_2018_8R9d_9kgg. I worked in close collaboration with prof. Marco Perugini from my department
Tom Coupé	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I was part of a team at the University of Canterbury who were involved in replications/reproductions of approximately 40 studies. We were involved in multiple aspects of the project, including data preparation, data analysis, and writing up of final reports for each study.
Jamie Cummins	0	0	1	0	1	1	0	0	1	0	1	0	0	0	I reviewed several replication preregistrations. In addition, I completed a replication of Woltin et al., 2011.
Aneta Czernatowicz-Kukuczka	0	0	1	0	1	1	0	0	0	0	0	0	0	1	I participated in planning the experiment, performed data analysis, and wrote the data description. I contributed to one experimental design process and analyzed one dataset. My work was conducted in collaboration with other team member involved in the study (Ewa Szumowska)
Joshua de Leeuw	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I conducted 3 replication studies (Rich_JournExPsychGen_2018_LbEB; Griffiths_JournExPsychGen_2011_J7ek; van Dijck_Cognition_2009_zN22). I worked with three of my students (Tessa Charles, Miles Bader, Rachel Ostrowski).

David Dobolyi	0	0	1	1	1	0	0	0	0	0	0	0	0	0	Led the Johnson et al. (2012) replication (https://osf.io/8vn2z/), including: 1) serving as the lead project PI on this replication in collaboration with the SCORE team (e.g., Zach Loomas); 2) the IRB submission and study approval (along with multi-site reliance agreements, etc.); 3) research/design of the Qualtrics surveys, consent/recruitment materials, and preparation notes; 4) preparation/cleaning of the merged (two-part) study data sets and data dictionary; and finally 5) programming of the study analysis script and generating results. Note: this replication required collaboration with several others for funding and data acquisition, including Ahmed Abbasi, Ryan Wright, and several other co-PIs outlined in the personnel section of the IRB form (including UVA and non-UVA sub-investigators): https://osf.io/cf8x9
James N. Druckman	0	0	1	0	1	0	0	0	0	0	0	0	0	0	I designed, oversaw implementation, and contributed to the analysis and write-up of the Tomz-van Houweling replication.
Jianhua Duan	0	0	1	0	1	0	0	0	0	0	0	0	0	0	I co-led the team at the University of Canterbury who were involved in replications/reproductions of approximately 40 studies. We were involved in multiple aspects of the project, including data preparation, data analysis, and writing up of final reports for each study.
Marin Dujmović	0	0	1	0	1	0	0	0	0	0	1	1	0	1	I was in charge of participant reimbursement and data collection, alongside Peter Allen I conducted the statistical analysis and contributed to the write-up of the SCORE report for the replication study (Project ID: Srinivasan_Cognition_2010_GXEW – Allen – Direct Replication - 67g8, OSF: https://osf.io/834de/)
Daniel J. Dunleavy	0	0	0	0	0	0	0	0	0	0	1	0	0	0	I participated as a reviewer, but honestly it was so long ago, I can't recall the specific studies.
Patrick K. Durkee	0	0	1	0	0	1	0	0	1	0	0	0	0	0	I conducted a pre registered secondary analysis using data published in other articles.
Cécile Emery	0	0	1	0	1	1	0	0	0	0	0	0	0	0	Direct réplication Crossley 2009 Journal of Organisation Behaviour, 5Xaw
Kevin M. Esterling	0	0	0	0	0	0	0	0	0	1	1	0	0	0	I served as editor for several papers, but I did not keep track of how many, and I was unable to recover the number from a search of my emails. I think Olivia can look up how many I edited and which ones.
Thomas R. Evans	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I planned and conducted one replication project (Lefevre_EurJournPersonality_2013_GNjz) in its entirety.
Anna Fedor	0	0	1	0	0	0	0	0	0	0	0	0	1	0	Planned and conducted a replication: I was a data analyst on the replication of a research claim from Malik et al. (2020)'s paper. Data finders were Sara Capitán and Tabaré Capitán. I also served as a reviewer for 2 protocols (Liu et al 2015, Yang et al, 2013).
Belén Fernández-Castilla	0	0	1	0	0	0	0	0	1	0	1	0	1	1	I am a member of the Replication team (co-authors are Ramljak and Santos) for a replication project named "Maluch_LearnInst_2015_5Awm". In the replication project, we analyzed the data collected by the data finder and completed the report, providing the code of the analysis conducted in R.
Nathan Fiala	1	0	1	0	1	0	0	0	0	0	1	0	1	1	I led PhD students in the replication of three papers, including conceptualization, analysis and writing.
James G. Field	0	0	1	0	1	1	0	0	1	0	0	0	0	1	I conducted three replication studies.
Nathan Fong	0	0	0	0	1	0	0	0	0	0	0	0	0	0	Collaborated with Sangsuk Yoon to collect replication data for Zhang (2009). Prepared/revised protocol, debugged software used to run experiment, recruited participants, analyzed and summarized results (for this experiment, not the overall study).
Miguel A. Fonseca	0	0	0	0	1	0	0	0	0	0	0	0	0	1	* Planned and conducted 3 replication projects (Goerg_JournLabEco_2010_WLpV, Goeree_Econometrica_2011_1574, Altmann_JournLabEco_2012_WLkV). * Served as an editor for peer review of replication designs (do not have the exact number or project codes handy). * Served as a reviewer for peer review of replication designs (do not have the exact number or project codes handy).
Alexandra L.J. Freeman	1	0	0	1	1	0	0	0	0	0	0	0	0	0	I oversaw a team of four (Dryhurst, Schneider, Roosenbeek & Kerr) undertaking 5 covid replication studies (Bertin, Pennycook, Rothgerber, Stanley & Wise), and before that, one non-covid (with Lawrence, replicating Yang). I helped design and write up each replication to SCORE standards.
Jeremy Freese	0	0	0	0	1	0	0	0	0	0	0	0	0	0	I am co-PI of TESS (Time-Sharing Experiments in the Social Sciences), which was used for one of the replications -- Tomz/van Howelling. Also, if relevant, I think I was on a supervisory committee that was tasked with being available if disputes arose in the context of a replication effort in Sociology.

Sandra J. Geiger	0	0	0	0	0	0	0	0	0	0	1	0	0	1	I served as a reviewer for 2 replication submissions. (I don't fully remember how many submissions it was but I believe it was 2.)
Jing Geng	0	0	0	0	1	1	0	0	1	0	0	0	0	0	I collaborated on data analytics to replicate research claims made by Anderson (2011) and Desmond (2015). My responsibilities included coordinating with the team, applying for IRB approval, and performing data cleaning and programming using R.
Laura M. Getz	0	0	1	0	1	1	0	0	0	0	0	0	0	0	I was involved in the replication of Lowe and Haws (2017) Study 1. This replication was completed in collaboration with Erin Hannon, Jared Leslie, and Mary Sanchez at the University of Nevada, Las Vegas: https://osf.io/dk3xz/.
Linda Marjoleine Geven	0	0	0	0	0	0	0	0	0	0	1	0	0	0	I served a reviewer for 1 replication submission.
Ilka Helene Gleibs	0	0	1	0	1	1	0	0	1	0	1	0	0	0	I served as a reviewer for 3 replication submissions (Review for Bersani and Doherty (2013), Ma-Kellams & Blascovich, 2011; Goh et al. (2020) COVID replication) and I also conducted a direct replication myself (Alves_92g)
Donna Pamella Gonzales	0	0	0	0	1	0	0	0	0	0	0	0	0	1	I prepared the data according to the specifications of the replication submission I handled on Fragile Families.
Janaki Gooty	0	0	1	1	1	1	0	0	0	1	0	0	0	1	I led two replications - Dumas et al and Popper et al. I worked with SCORE staff and graduate students to design the protocols, collect data, analyze it and write the report.
Amélie Gourdon-Kanhukamwe	0	0	0	0	0	0	0	0	0	1	1	0	0	1	According to my working hours records, I served as reviewer for 11 submissions and as editor for 22 submissions, although screening back payment agreements and Gdocs I have once worked on, I can confirm only 27 names (7 reviews and 20 editing jobs). Of these, 26 were replications: the list of identified studies is at https://ameliegourdonkanhukamwe.notion.site/2fd4b161b8994bb39d75cb097ece5f22?v=1faf926789d74544be1bd377c2330d0e&pvs=4
Cristina Grecules cu	0	0	0	0	1	1	0	0	0	0	0	0	0	1	I contributed to one study, namely the replication of "The weekend matters: Relationships between stress recovery and affective experiences" (Fritz et al., 2010). I participated in conducting the k17 replication study by identifying institutions eligible for the study, contacting the institutions, and recruiting kindergarten teachers (Investigation). I also took part in the development and adjustment of the recruitment strategy (Methodology). Furthermore, I helped review and edit the final report submitted for this replication study.
Siobhán M. Griffin	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I led the Irish team to replicate the findings of Fritz et al 2010, with my collaborators in Ireland - Cillian McHugh and Ellie Shackleton. This involved acquiring ethics from BRANY, contacting childcare facilities across Ireland to recruit participants and follow them across three time points. Included creation of the survey on qualtrics, data collection, data cleaning, data collating, analysis and write up. There was also a German team collecting German data on this. We worked with Zach and Bea to manage the project.
Lusine Grigoryan	0	0	0	0	1	1	0	0	0	1	0	0	0	1	Together with Vladimir Ponizovskiy, I supervised the team that conducted the k17 replication study in Germany. I was involved in all stages of the project: from ethics approval, preregistration, and methods adaptation, to data collection and preparation of the final report.
Martina Grunow	0	0	0	0	1	0	0	0	0	0	0	0	0	0	Collecting and preparing data according to certain claims for two papers
Nicholas Gunby	0	0	1	0	0	0	0	0	0	0	0	0	0	0	I wrote code to replicate the Baxter et. all Social Forces study - collaborated closely with Bob Reed and Jane Duan
Braeden Hall	0	0	1	0	1	1	0	0	0	0	0	0	0	0	I designed the replication study for our participating lab, carried out data collection, and performed the replicated analyses. And, I look forward to contributing meaningfully to the paper.
Paul H. P. Hanel	0	0	1	0	1	1	0	0	1	0	0	0	0	1	I led one of the replication studies. Specifically, I contributed to setting up the study, collected and analysed the data, and wrote a report.
Erin E. Hannon	0	0	1	0	1	0	0	0	0	1	0	0	0	0	We replicated a social psychology/marketing study about cross-modal perception in our lab. For this project I directed two research assistants who helped me prepare the experiment interface, collect the data from human subjects, format the data for sharing, and run statistics on the data.
Sam Harper	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I collaborated with Arijit Nandi to replicate the findings from Table 3 reported in Ishida et al. (Soc Sci Med. 2010 Nov;71(9):1653–61) that showed an association between exposure to intimate partner violence and mental health in women from Paraguay. We attempted to replicate this result with similar data from Nicaragua. Our replication results were similar in both direction and magnitude of the coefficient reported in the original analysis by Ishida et al.

Marco Jürgen Held	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I did a replication study of a research claim from Kang et al. (2009), from Journal of Experimental Social Psychology
Louis Hickman	0	0	0	0	1	1	0	0	0	0	0	0	0	0	I designed, recruited participants for, conducted, and also reviewed and advised on the formal analysis of the replication of Gabriel_JournAppPsych_2011_egX9.
Nathan C. Higgins	0	0	1	0	1	1	0	0	1	0	0	1	0	1	I provided the experiment presentation code, conducted controls to verify temporal precision of stimulus presentation, and developed Matlab routines to replicate the original data analysis and generate figures of results.
Svenja Hippel	0	0	1	0	1	0	0	0	1	0	0	0	0	0	I conducted a "bushel" direct replication, together with Sven Hoeppner and Heinrich Nax. The SCORE Study ID is: Hoeppner_Rodriguez_Lara_23m12
Sven Hoeppner	0	0	1	0	1	0	0	0	1	0	0	0	0	0	I conducted one "bushel" direct replication study, together with Svenja Hippel and Heinrich Nax. Score Study ID is: Hoeppner_Rodriguez_Lara_23m12
Sanghyun Hong	0	0	1	0	1	0	0	0	0	0	0	0	0	0	I was part of a team at the University of Canterbury who were involved in replications/reproductions of approximately 40 studies. We were involved in multiple aspects of the project, including data preparation, data analysis, and writing up of final reports for each study.
Thomas J. Hostler	0	0	1	0	1	1	0	0	1	0	0	0	0	0	Conducted one replication study (Vollmann et al, x3KP) including design of replication study, collection of data, and analysis of data. This with done with Sofia Persson (Leeds Beckett University)
Michael Inzlicht	0	0	0	0	0	0	0	0	0	1	0	0	0	0	My lab replicated Yang, Vosgerau, & Loewenstein, 2013. My primary role was in supervising and providing material support for Hause Lin, who ran the study in my lab and who was a PhD student with me until recently (when he graduated).
Kamil Izydorczak	0	0	0	0	0	0	0	0	0	0	1	0	0	0	I reviewed replication projects for two studies: de Langhe_JournConsRes_2014_RJb - Walasek_6m4k, Fox_JournLabEco_2009_LyLd_944y (Bailey/Reed)
Bastian Jaeger	0	0	1	0	1	0	0	0	0	0	0	0	0	0	I conducted a replication (Yang_JournMarketRes_2013_G1Lr - Jaeger - New Data Replication - 387)
Kristin Jankowsky	0	0	1	0	1	1	0	0	1	0	0	0	0	1	
Johannes Jarke-Neuert	0	0	1	0	1	1	0	0	1	0	0	0	0	1	I conducted one entire replication study with new data (Janssen_ExpEco_2011_VvZX - Jarke-Neuert - New Data Replication - 4182), analyzed the data, reported the findings and reviewed the reports of several other studies.
Matthew Jensen	0	0	0	0	1	0	0	0	0	0	0	0	0	0	Helped with data collection for the Johnson et al (2012) replication (https://osf.io/8vn2z/)
Biljana Jokić	0	0	0	0	1	1	0	0	0	0	1	0	0	1	I was a member of the research team on conducting the replication study: Direct replication of a research claim from Luttrell et al, 2016, in the Journal of Experimental Social Psychology
Daniel Jolles	0	0	0	0	1	1	0	0	1	0	0	0	0	0	I was a member of the research team with Prof. Marie Juanchich for the replications of King & Bryant, 2017 and Ku, 2014.
Phillip Jolly	0	0	0	0	0	1	0	0	0	0	1	0	0	0	I served as a reviewer for 1 replication submission, and designed a replication study for Liu et al. (2015) that went through preregistration review and was approved, but that was unable to be conducted because of COVID-19. I am not sure that warrants coauthorship given that the research was not able to be conducted because participants had to be recruited in-person in hotels, and hotels shut down for most of 2020. That replication preregistration was conducted with Anna Mattila at Penn State.
Angela M. Jones	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I conducted one replication study with Dr. Sean Patrick Roche (also from Texas State). The study we replicated was Pickett, J. T., & Baker, T. (2014). The pragmatic American: Empirical reality or methodological artifact?. Criminology, 52(2), 195-222.
Marie Juanchich	0	0	1	0	1	1	0	0	1	0	0	0	0	0	We replicated a study, analysed the data and reported on the findings
Pavol Kačmár	0	0	1	0	0	0	0	0	1	0	0	0	0	1	I have conducted DAR (SCORE study id: Vadillo_247z3).

Hansika Kapoor	0	0	1	0	0	0	0	0	0	0	1	0	0	0	Reviewer for 23 papers: Stanley_covid_b3G4_98g Pennycook_covid_7NEL_9k1y Malik_covid_Y3jx_348 Pfattheicher_covid_yZD4_y006 Du_covid_2NAG_41z0 Bohnke_EurSocioRev_2017_xGGO_y11 O'Brien_AmSocioRev_2015_7X54_93k7 Denson_AmEduResJourn_2009_zb3Y_41k2 Du_covid_2NAG_41z0 Niehaus_AmEduResJourn_2014_BlRQ_546 Montez_Demography_2014_3aPw_05g8 Kim_CompEdu_2014_YWep_75g6 Past√∂tter_Cognition_2013_EQxa_3z3k Seaton_AmEduResJourn_2010_Blxd_6778 Berg_covid_qKPb_k127 Baxter_SocialForces_2015_z0v1_y410 Kausel_OrgBehavior_2015_5XEE Petit_JournBusRes_2017_9R9X Griffiths_JournExPsychGen_2011_J7ek Bhattacharjee_JournPerSocPsy_2017_Br0x King_JournOrgBehavior_2017_Q1dl Raley_JournMarFam_2012_D2LY Weidmann_2g7ky Montez_2y4gm Karraker_2g79y Anderson_329k Li_6m34m van Gastel_21487 Liang_23g12 BATESON_2k5g2 Andrews_95my. Replication researcher for Gerhold
Andjela Keljanovic	0	0	0	0	1	0	0	0	0	0	0	0	0	0	Data collection for Replication of Research from Luttrell et al. (2016), The Journal of Experimental Social Psychology Replication Team: Iris Žeželj, Biljana Jokić, Danka Purić, Bojana Bodroža
Samjhana Koirala	0	0	1	0	1	1	0	0	1	0	0	0	1	0	Replication of "On the interpretation of bribery in a laboratory corruption game: moral frames and social norms" by Ritwik Banerjee
Marta Kołczyńska	0	0	1	0	1	0	0	0	0	0	0	0	0	0	I worked on 4 replication submissions as analyst.
Dimitra Kouroupaki	0	0	1	0	0	0	0	0	0	0	0	0	0	0	I coded data for one replication study (original Torelli and Shavitt), phase one and two.
Ulrich Kühnen	0	0	0	0	1	0	0	0	0	0	0	0	0	0	-Participated in conducting the k17 replication study by identifying institutions eligible for the study, contacting the institutions, and recruiting kindergarten teachers. -Reviewing and editing the final report of the k17 replication study
Michelangelo Landgrave	0	0	1	0	1	0	0	0	0	0	0	0	0	0	I conducted and helped design the Einstein and Glick (2017) replication.
Michael J. Larson	0	0	0	0	1	0	0	0	0	0	0	0	0	0	My colleagues and I conducted data collection for one of the replication submissions.
Lyonel Laulié	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I lead a replication project (get IRB approval, pre-registered the study, collected data, maintain relationships with the organization where data was conducted, analyzed data, produced a summary of results, prepared documents for sharing results in COS)
Alice C E Lawrence	0	0	1	0	1	0	1	0	0	0	0	0	0	0	I collected the data, including setting up the online survey, undertook the formal replication analysis and subsequently wrote a report for submission to SCORE for one replication. Also liaised with the SCORE team.
Joel M. Le Forestier	0	0	1	0	1	1	0	0	0	0	0	0	0	1	I planned and conducted 1 replication study.
Katelin E. Leahy	1	0	1	0	1	1	0	0	0	1	1	0	1	1	For our replication project (Nelson), I (with help from students) programmed the survey and edited/spliced/hosted the videos for the study. I coordinated data collection and training of RAs, and supervised data analysis/reporting, with my fellow graduate students Mariah Purol and Jeewon Oh. I reviewed several replication/pre-registrations. My fellow graduate students and I wrote the pre-registration, conducted the analysis, wrote up the results, and prepared the final report and data/syntax files for review.
Sungmok Lee	1	0	0	0	1	0	0	0	0	0	0	0	1	1	I conducted data collection and served as a writer and reviewer for 1 replication submission.
Jared Leslie	0	0	1	0	1	0	0	0	0	0	0	0	0	0	Assisted in running one replication submission and analyzing/reporting data for that submission. I worked on the Lowe_Journ_Market_Res_VGae study and assisted with project set-up, analysis, write-up. My collaboration on the project was directly through my PI (Dr. Erin Hannon) and fellow collaborator (Dr. Laura Getz).

Savannah C. Lewis	0	0	1	0	1	1	0	0	1	0	0	0	0	1	During the SCORE project, I created a survey link for participants to use by coding a formr link. I also uploaded and monitored the Mturk participants for our replication. Next role I played was analysis and coding the data to be sent to the lead team and will also assist in the reviewing and editing process of the manuscript.
Christopher Limnios	0	0	1	0	1	1	0	0	1	0	1	0	0	0	Wrote scripting code to replicate and validate statistical claims made by authors in original paper. Validated statistical claims made by authors in two separate original papers.
Hause Lin	0	0	0	0	1	1	0	0	1	0	0	0	0	0	I conducted the replication study for Yang et al. (2013), wrote the analysis code, and analyzed the data.
An-Chiao Liu	0	0	1	0	1	1	0	0	1	0	0	0	0	0	
John Wills Lloyd	0	0	0	0	0	0	0	0	0	1	1	0	0	0	I was pleased to have opportunities to guide multiple teams as they prepared plans to conduct replications.
Elliot A Ludvig	0	0	0	0	1	1	0	0	0	0	0	0	0	0	I was part of the team that replicated the De Langhe et al. (2014). I helped with the design of the study, getting ethics approval, writing and adjusting the pre-registration, reviewing study materials, and composing the final report. There were 3 other members of our team: Lukasz Walasek, Gordon Brown, and Finnian Wort.
Dermot Lynott	0	0	0	0	0	0	0	0	0	0	0	0	0	1	I served as a reviewer for 4 replication submissions. Those studies were: Rinaldi_Cognition_2016, mason_journexpsychgen_2012, pastötter_cognition_2013, vermeulen_eurjournpersonality_2010.
Jordan MacDonald	0	1	1	0	1	1	0	0	1	0	1	1	1	1	Prepared all materials for the replication of a paper by Yiend and Colleagues (2015) on selective attention in generalized anxiety disorder. Role included: creating attentional task in Python, writing some of the intro, the entire methods, results, and conclusions/discussion, running statistical analyses.
Peter Mallik	0	0	1	0	1	0	0	0	0	0	0	0	0	0	I was involved in 4 separate SCORE project initiatives. A large part of the work done at the beginning of the endeavor impacted future replication attempts. I was a PI on Bursztyn, Pfattheicher, Netemeyer, and Zunick replications (all registered on OSF). ID numbers (2g7z2, 286, 46z8, and wz9).
Daniel J. Mallinson	0	0	1	0	1	1	0	0	1	0	1	0	0	0	I served as a replication researcher for one paper and pre-registration reviewer for three or four studies. I am unsure of exactly how many as reviewer.
Daniele Marinazzo	0	0	0	0	0	0	0	0	0	0	1	0	0	0	I participated in the review of five pre registration claims
Corinna S. Martarelli	0	0	0	0	0	0	0	0	0	0	1	0	0	0	I served as a reviewer for 3 replication submissions (Bosma et al., 2017; Kieffer, 2011; Vadillo et al., 2016)
Joshua Matacotta	0	0	0	0	0	0	0	0	0	0	1	1	1	1	I served as an editor for peer review of replication designs. I served as a reviewer for peer review of replication designs. Worked with Bri, Rich, Bob and others. There were a few. For example, SCORE Preregistration review for a replication of a claim from Šrol et al. (2020); González-Marrón & Martínez-Sánchez (2020); ..
Andrew McBride	0	0	1	0	1	0	0	0	1	0	0	0	0	0	I, along with Dr. Janaki Gooty, conducted 1 replication study. This involved planning, creating surveys, collecting data, and analyzing the data.
Cillian McHugh	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I was involved in a replication of research claim from (Fritz) et al. (2010) in an Irish context with collaborator Siobhán Griffin (local team leader). I supported the team leader in securing ethics through Brany, and contacting childcare facilities to collect data. I ran the analysis.
Gail McMillan	0	0	1	0	1	1	0	0	0	0	0	0	0	0	I was a postdoctoral fellow under supervision of Dr Marina Milyavskaya assigned to replicate Kappes et al.(2012) (RR id: g9mm). I was responsible for preparing the pre-registration, setting up the experiment within software, recruitment/management of participants, formal analysis of the data, and preparation of the final report.
Esteban Méndez	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
Mitchell Metzger	0	0	0	0	1	0	0	0	0	0	0	0	0	0	I completed a replication (Netemeyet et al. 2017)
Michalis P. Michaelides	0	0	1	0	1	1	0	0	0	0	0	0	0	0	I prepared data from existing sources and documentation for 2 replication submissions: Zimmerman and Vukovic papers. For the former, along we the colleague who conducted the analysis, we published it as a replication paper (Michaelides & Durkee, 2021).
Johannes Michalak	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
Leticia Micheli	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I participated in the replication of one research claim from Ihme & Tausendpfund (2018). Together with my team, we were responssible for designing the replication, collecting data and anaysing results.
Jeremy K. Miller	0	0	0	0	0	0	0	0	0	0	1	0	0	1	I reviewed several reviews for replication submissions, for example that of Savani et al. 2010.

Marina Milyavskaya	0	0	1	0	1	1	0	0	0	1	0	0	0	1	Together with my postdoctoral fellow Gail McMillan (whom I supervised), we conducted a replication study of Kappes et al., 2012 (project ID: Kappes_JournExpSocPsych_2012_9J1 - Direct Replication – g9mm).
Daniel C. Molden	0	0	1	0	1	1	0	0	1	0	0	0	0	1	I conducted the replication study: Sandra_Cognition_2018_0qar - Molden - Direct Replication - 20g
Ambar G. Monjaras	0	0	1	0	1	0	0	0	0	0	0	0	0	0	I completed 54 experiments for 28 subjects (2 experiments per subject) and collected data for each of these experiments. I analyzed data of 26 subjects that were eligible within the inclusion criteria. I collaborated closely with Dr. Joel Snyder, PhD, Dr. Nathan Higgins, PhD, and Mary Sanchez.
David Moreau	0	0	0	0	1	0	0	0	0	0	1	0	0	0	Edited replication submissions and served as a reviewer on replication submissions.
Audrey Morrow	0	0	1	0	1	0	0	0	1	1	1	1	1	1	I was the lead graduate student on the assigned replication study and was responsible for training two undergraduate research assistants on data collection. I was primarily responsible for programming and carrying out the main analysis, creating two main figures to visualize the data, and writing an original draft of the replication summary. Finally, I assisted in conducting two exploratory analyses and revising the replication summary.
Cristóbal Moya	0	0	1	0	0	1	0	0	1	0	0	0	1	1	I conducted one replication with secondary data as part of the Center for Open Science effort in the SCORE project, developing the design, conducting the study and reporting the results according to the project guidelines. Also, I worked doing claim extractions for multiple papers in the SCORE program
Liad Mudrik	0	0	0	0	0	0	0	0	0	0	1	0	0	0	I reviewed 1 replication submission, of the study by Linkenauger et al., 2009
Laetitia B. Mulder	0	0	1	0	1	1	0	0	0	0	0	0	0	0	Planned and conducted a replication. I worked closely with Susanne Tauber in this task.
Katie A. Munt	0	0	1	0	1	0	0	0	0	0	0	0	0	0	In relation to 'Replication of a research claim from Rubin et al. (2010) from Leadership Quarterly', together with Niklas Steffens and Kim Peters, I assisted with research design, pre-registration on OSF, creation of study materials, data collection, data analysis, and preparation of documents for the SCORE team.
Arijit Nandi	0	0	1	0	0	1	0	0	0	0	0	0	0	0	Collaborated with Sam Harper to replicate a finding from Ishida et al. showing an association between exposure to intimate partner violence and mental health in women from Paraguay.
Kathryn Nason	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
Carolin Nast	0	0	0	0	1	0	0	0	0	0	0	0	0	0	I worked on two phases of the score project. First, I conducted several Data Analytic Replications (I am not sure if I managed to include all in my list here): (1) Teney (Paper ID qXX2), (2) Pana (Paper ID p2qE), (3) Hossain (Paper ID b344), (4) Chen (Paper ID yAPR), (5) Gonzalez Marron (Paper ID W3vN), (6) Seaton (Paper ID Blxd). In the next phase, I worked on (1) the O'Brien (Paper ID 7X54) Bushel DAR, Bushel ADR and Single ADR; (2) the Maluch (Paper ID 5Awm) Single SDR; (3) the Seaton (Paper ID Blxd) Bushel SDR and Bushel DAR. I collaborated mostly with Marco Ramljak (first and second phases), and also partly with Ferdinand Wintermantel (only partly in the second phase).
Gideon Nave	0	0	0	0	1	0	0	0	0	0	0	0	0	0	My lab performed one replication study
Heinrich H. Nax	0	0	1	0	1	0	0	0	0	0	1	0	1	1	* Conducted a "bushel" direct replication, together with Svenja Hippel and Sven Hoeppner (SCORE Study ID is: Hoeppner_Rodriguez_Lara_23m123). * Served as a reviewer for peer review of replication projects underlying “Analyst interpretation” * Suggested hypotheses and interpretations of results
Florian Neubauer	1	0	1	0	1	0	0	0	0	0	0	0	1	1	I designed and conducted a replication of the following paper: Banerjee, R. (2016). On the interpretation of bribery in a laboratory corruption game: Moral frames and social norms. Experimental Economics, 19(1), 240–267. https://doi.org/10.1007/s10683-015-9436-1 I was involved in designing the replication and survey instruments, collecting the data, running the formal analysis, and writing the paper.
Phuong Linh L. Nguyen	0	0	1	0	0	0	0	0	0	0	0	0	0	0	I conducted formal analysis of a claim in a published paper.
Austin Lee Nichols	0	0	0	0	0	1	0	0	0	0	1	0	0	0	I served as a reviewer for Chittoor 2009 project

Gustav Nilsonne	0	0	0	0	1	0	0	0	0	1	1	0	0	0	I served as editor and reviewer. I also did some data finding and preparation, but am not sure if any of those were used for the replication part of the SCORE project of only for the reproduction part . I did not keep records of how many studies I reviewed, but have now attempted to check by going through the reimbursement forms. I have found 1 study where I was listed as editor and 11 studies where I was listed as reviewer. There are a further 12 studies for which I have received reimbursement but where the form did not specify my role.
Ernest O'Boyle	0	0	0	0	1	0	0	0	0	0	0	0	0	0	I served as an AE on the SCORE project. I oversaw the review process, reconciled reviewer and author differences, and ultimately accepted or rejected the manuscript.
Jule Oettinghaus	0	0	0	0	1	0	0	0	0	0	0	0	0	0	I participated in conducting the k17 replication study by identifying institutions eligible for the study, contacting the institutions, and recruiting kindergarten teachers. Furthermore I helped reviewing and editing the final report of the k17 replication study.
Jeewon Oh	0	0	1	0	1	1	0	0	0	0	1	0	1	1	Our research team included William Chopik, Katelin Leahy, and Mariah Purol. We replicated an existing study and collected new data to test this question. I helped coordinate date collection and analyzed data with my team members.
Adoril Oshana	0	0	1	0	1	0	0	0	1	0	1	1	1	0	I executed the replication study with my adviser Dr. George C. Banks. I secured IRB approval, coordinated participant recruitment, curated and analyzed the data, drafted the original manuscript, and created visualizations. Dr. Banks provided background context, overall guidance, and managed communications, while I handled the core investigative, analytical, and drafting responsibilities.
Thomas Ostermann	0	0	0	0	1	0	0	0	0	0	0	0	0	1	
Rachel P. Ostrowski	0	0	0	0	1	1	0	0	1	0	0	0	0	0	I worked alongside Josh de Leeuw and Miles Bader. During the summer of 2021, we performed a replication study of Rich & Gureckis (2018) and a generalizability study of Griffiths & Tenenbaum (2011). Miles and I contributed equally as undergraduate researchers; Josh was our PI. We worked together on most every step of the replication and generalization processes, from experiment design (when applicable), to the online experiments' software implementation, through data collection and analysis. The extent of my contribution to the overall project is limited to these two studies -- I'll defer to Josh's submission if there happen to be any inconsistencies regarding our specific contributions.
Abiola Oyebanjo	0	0	1	0	0	0	0	0	0	0	0	0	0	0	I led the data analysis for a replication study.
Radoslaw Panczak	0	0	1	0	0	0	0	0	0	0	0	0	0	1	I worked as analyst on two papers (Fielding-Miller_covid_R3pV_615k and Kim_SocSciMed_2016_AqDO_65z92_SDR). I reviewed and edited final manuscript.
Jamie Patrianakos	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I performed 1 replication study and report with guidance from Robyn Mallett.
Ignacio Pavez	0	0	0	0	1	1	0	0	0	0	0	0	0	0	I was part of the research team who conducted the replication study: Becker_JournOrgBehavior_2015_J40k - Laulié - Direct Replication - 67z8. I collaborated closely with Prof. Lyonel Laulié (principal researcher), and Gabriel Briceño. My role was co-researcher.
Yuri G. Pavlov	0	0	1	0	1	1	0	0	1	0	0	0	0	1	Conducted a replication of one study: Hurst et al. (2013), https://osf.io/hz97d/
Sofia Persson	0	0	1	0	1	1	0	0	0	0	0	0	0	0	I collaborated with Dr Thomas Hostler (Manchester Metropolitan University) to carry out 1 replication study.
Marco Perugini	0	0	1	0	1	1	0	0	0	0	0	0	0	0	I conducted and analyzed 1 replication, Christensen et al. (2018) published in the European Journal of Personality, project code Christensen_EurJournPersonality_2018_8R9d_9kgg. I worked in close collaboration with prof. Giulio Costantini from my department
Kim Peters	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I ran Peters_Crossley_69y36 , Thomas_JournPerSocPsy_2014_xvrb – Peters – z591 (replication) and contributed to Rubin_LeadQuart_2010_lJ0w_931g (replication)

Constant Pieters	0	0	1	0	0	0	0	0	1	0	0	0	0	0	Niels van de Ven collected replication data and invited me (Constant Pieters) to analyze those. I inspected the original article to infer the analysis method, analyzed the replication data, and wrote the analysis code; all with continuous input from Niels van de Ven.
Vladimir Ponizovskiy	0	0	1	0	1	1	0	0	1	1	0	0	0	0	I coordinated one of the replication projects (k17), supervised the administrative side of the project, participated in the adaptation of the study materials, adapting the methods of the original study for replication, supervised and participated in collecting data from German kindergartens, maintained the project's OSF directory and was the lead author on the replication report.
Nathaniel D. Porter	0	0	1	0	1	1	0	0	1	1	1	0	0	0	I performed 4 replication studies, 2 new data replications with data collection of Raver & Nishii 2010 with Cozette Comer, data finding and replication for Anderson 2011 with Anirudh Tagat & Jing Geng, and data finding and replication for Desmond 2015 with Jing Geng. I additionally performed data finding for Noah 2018. I served as preregistration review editor for 25 replication studies in my role as lead sociology review editor (see spreadsheet in comments for studies), as well as reviewing 5 other replication studies where I was not an editor.
Jason M. Prenoveau	0	0	0	0	0	0	0	0	0	0	1	0	0	0	I was a member of the preregistration review team. In this capacity, I reviewed preregistration materials for the replication studies that were a part of SCORE. This happened in late 2019 and I do not remember how many studies I served as a reviewer for (maybe 1 or 2).
Danka Purić	0	0	1	0	1	1	0	0	0	0	0	0	0	0	With my colleagues, I conducted the direct replication of a research claim from Luttrell et al. (2016), from The Journal of Experimental Social Psychology. We collected the data, performed the required statistical analyses and wrote a report on the findings.
Mariah F. Purol	0	0	1	0	1	0	0	0	1	0	1	0	1	1	I created materials, collected data, analyzed data, and wrote up results for 1 replication submission.
Arathy Puthillam	0	0	1	0	1	0	0	0	1	0	0	0	0	0	
Kimberly A. Quinn	0	0	0	0	0	0	0	0	0	1	1	0	0	1	
Marco Ramljak	0	0	1	0	1	1	0	0	1	0	0	1	0	0	I collaborated closely in all projects with Carolin Nast and Ferdinand Wintermantel. We conducted multiple replication projects and were involved in phase 1 and 2 of the overall projects.
W. Robert Reed	0	0	1	0	1	0	0	0	0	1	0	0	0	0	I co-led a team of 10 researchers, mostly based at the University of Canterbury, who were performed replications/reproductions of approximately 40 studies. We were involved in multiple aspects of the project, including data preparation, data analysis, and writing up of final reports for each study.
Michaela Ritchie	0	0	0	0	1	0	0	0	0	0	0	0	1	1	I played a role in data analysis and I contributed to the writing of all drafts of this manuscript.
Margaret Ritzau	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
Sean Patrick Roche	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I conducted 1 replication (PICKETT_Criminology_2014_P1rY_99kg). I collaborated closely with my colleague Angela Jones.
Romina Rodela	1	0	0	0	0	1	0	0	0	1	0	0	1	1	If I remember correctly have served as a reviewer for 1 submission
Jan Philipp Röer	0	0	1	1	1	1	0	0	1	1	1	0	0	1	I have planned and conducted a replication study together with Thomas Ostermann and Johannes Michalak (https://osf.io/apv5t/) and served as a reviewer for a couple of replication submissions. I also edited 20-30 submissions, but I haven't kept track of the exact number.
Ivan Ropovik	0	0	1	0	1	1	0	0	1	0	0	0	0	1	Me and my team were working on the following replication: Stahl_JournPerSocPsy_2012_gbl9 (replication 393)
Jacob Rothschild	0	0	1	0	1	0	0	0	0	0	0	0	0	0	
Justine Saal	0	0	0	0	1	1	0	0	0	0	0	0	0	0	Participated in conducting the k17 replication study by identifying institutions eligible for the study, contacting the institutions, and recruiting kindergarten teachers; discussions about recruitment strategies and materials used in the study
Hani Safadi	0	0	0	0	1	0	0	0	0	0	1	0	0	0	Helped with data collection for the Johnson et al (2012) replication (https://osf.io/8vn2z/) and validation of the accuracy of the statistical outcomes and interpretation of the findings.
Jason Samaha	1	0	1	1	1	1	0	0	1	1	0	1	0	1	Designed and oversaw research, contributes to data analysis and writing.

Mary Sanchez	0	0	0	0	1	1	0	0	1	0	0	0	0	0	I worked on two replication submissions: Replication of a Research Claim from Lowe and Haws(2017), & Petersen et al. (2017) Experiment 1. On both projects I worked with experiment creation, data collection, and aided in some analysis for both projects. On project one replication of claims from Lowe and Haws (2017) I worked closely with Jared Leslie under Erin Hannon, and project two replication of claims from Petersen et al. (2017) I worked closely with Ambar Monjaras and Nathan Higgins under Joel Snyder.
Soorya Sankaran	1	0	1	0	1	1	0	0	1	0	1	1	0	1	I helped with the planning, coding, data collection, and presentation of an earlier version of the replication study, and reviewed an early draft of the paper..
David Santos	0	0	1	0	0	0	0	0	1	0	1	0	1	1	I am part of the Replication team (co-authors are Ramljak and Fernández-Castilla) for a replication project named "Maluch_LearnInst_2015_5Awm" We analyzed the data collected by the data finder and completed the report, providing the code of the analysis conducted in R.
Amanda C. Sargent	0	0	1	1	1	0	0	0	0	0	0	0	1	0	I worked on two replication studies, one of which I conducted the full investigation from submitting the funding requests to SCORE and managing the approval process, to data collection and analysis and writing and submitting the final report (replication of Popper & Amit 2009). I served as a data analyst for the other replication (meta-analysis - Liu, 2016).
Marian Sauter	0	0	0	0	0	0	0	0	0	0	1	0	0	0	I reviewed around 3-4 replication submissions
Kathleen Schmidt	0	0	1	0	1	1	0	0	1	0	1	0	0	0	I completed two replication studies as the primary investigator: Seuntjens_JournPerSocPsy_2015_PNPz - Schmidt - Direct Replication - 5zg9 (Braeden Hall provided support) and Alves_PsychologSci_2018_AvOr - Schmidt - 24716 (bushel replication; Bruce Clark provided support). I also prepared another replication study but could not not execute it because DARPA wouldn't give it HRPO approval (Biernat, 2013). I reviewed at least 3 replication pre-registrations (Hurst et al., Karraker & DeLamater, and Gerhold et al.).
Landon Schnabel	0	0	0	0	1	1	0	0	1	0	1	0	0	1	I prepared a study, served as a reviewer for several replication submissions, and reviewed a draft of the paper.
Amber N Schroeder	1	0	1	1	1	1	0	0	0	1	1	0	1	1	
Sebastian W. Schuetz	0	0	0	0	1	0	0	0	0	0	0	0	0	0	Helped with data collection for the Johnson et al (2012) replication (https://osf.io/8vn2z/).
Brendan A. Schuetze	0	0	1	0	1	1	0	0	1	0	0	0	0	1	In collaboration with Veronica X. Yan, I carried out the planning, pre-registration, data collection, and analysis of the Muis and Franco (2009) replication study.
Michael Schulte-Mecklenbeck	0	0	1	0	0	1	0	0	0	0	0	0	0	0	I edited 2 submissions and reviewed 3 (?)
Astrid Schütz	0	0	0	0	1	0	0	0	0	1	0	0	0	1	I conducted and supervised a study
Eric L. Sevigny	0	0	1	0	1	1	0	0	1	1	1	0	0	1	I was the PI or Co-PI Replication Researcher on two projects (Weidmann_BritJournPoliSci_2013_JRpA-Sevigny_mk67; BERSANI_Criminology_2013_zmYY_g5m-Shakya/Sevigny). I was also the Action Editor for one replication project (BERSANI_Criminology_2013_zmYY - Adamkovic - 6zz3k). I supervised a graduate student in this work, and I actively reviewed, edited, and commented on the draft original manuscript.
Ellie Shackleton	0	0	0	0	1	0	0	0	0	0	0	0	0	0	I was part of the Irish team to replicate the findings of Fritz et al 2010, with collaborators Siobhán Griffin & Cillian McHugh. This involved contacting childcare facilities across Ireland to recruit participants and following them across three time points to collect data.
Richard M. Shafranek	0	0	0	0	1	1	0	0	0	0	1	0	0	0	Worked with James N. Druckman and Jeremy Freese on replication of a Tomz & Van Houweling 2009. Drafted materials including pre-registration plan and IRB protocol, responded to reviews, interfaced with third party vendors, assisted with methodological preparations.
Samuel Shaki	0	0	1	0	1	1	0	0	1	1	1	1	1	0	
Shishir Shakya	0	0	1	0	1	1	0	0	1	0	1	0	1	1	
Miroslav Sirota	0	0	1	0	1	1	0	0	1	0	0	0	0	1	A replication of a study that included pre-registration, materials preparation, data collection and analysis
Matthew Ryan Sisco	0	0	1	0	0	0	0	0	1	0	0	0	0	0	I prepared three replication datasets (https://osf.io/bj9ea/, https://osf.io/87qs6/, https://osf.io/bmwsv/)

Maksim M. Sitnikov	0	0	1	0	0	0	0	0	0	0	0	0	0	1	My role has mainly been assisting Robbie, Marjan, and Marcel in computing relevant variables for both the original and replication studies. In particular, besides being engaged in coding the types of effect sizes for the original set of studies, I was involved in editing and applying R scripts used for computing effect sizes and other variables of interest (e.g., statistical power). In addition, I was involved in communicating data-related requests to the OSF team (first to Melissa and then to Zach), particularly updating them about the types of data that were missing in the files with coded original/replication studies and that we needed to compute the variables of interest.
L. Robert Slevc	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I conducted and analyzed one replication study (of Lai & Poletiek, 2011, Cognition): https://osf.io/78fvn/
Laura Smalarz	0	0	1	0	1	1	0	0	0	0	1	0	0	0	I preregistered and carried out a study to test a claim from Biernat and Sesko (2013). I analyzed the data and wrote the final SCORE report for the replication.
Colin Tucker Smith	0	0	0	0	1	0	0	0	0	1	0	0	0	1	I worked with one of my graduate students to conduct a data collection (including data analysis and write-up).
Joel S. Snyder	0	0	1	0	1	1	0	0	1	1	0	0	0	1	I worked with and supervised 2 post-baccalaureate students and 1 post-doctoral fellow in my lab to set up, run, analyze, and write up 1 of the replicated experiments. The set up including programming the experiment anew in our lab's Presentation software and purchasing and integrating new pieces of equipment into our lab to run the study.
Nicolas Sommet	0	0	1	0	1	1	0	0	1	0	0	0	1	1	My co-author and I undertook a replication study utilizing secondary data. This involved locating the appropriate data, managing the data management process, analyzing the results, and compiling a comprehensive report.
Fatih Sonmez	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I managed the "Ku_JournEnvPsych_2014_YpZZ - RRTeam_Sönmez_ML - Direct Replication - 8y1" project. I preregistered the replication, collected the data, prepared the scripts, analyzed the data, and reported the findings.
Barbara A. Spellman	0	0	0	0	0	0	0	0	0	1	1	0	0	0	SCORE Editorial Team Action Editor worked on (according to e-mail): Linkenauger_PsychologSci_2009_7WjP; Kappes_JournExpSocPsych_2012_9J1; Pickett et al. (2014) -- so 3 submissions, 89 e-mails; I found a few potentially cross-cutting issues that generated changes to what we ask for from pre-registration protocols.
Natalia Stanulewicz-Buckley	0	0	0	0	0	0	0	0	0	0	1	0	0	1	The role I undertook was to review replication protocols and liase with authors to achieve a satisfactory protocol at the end. I have done that for multiple protocols but don't remember the exact number. I edited/reviewed the draft too.
George Stock	0	0	0	0	1	0	0	0	0	0	0	0	0	0	I coded 350 effect sizes for a creativity meta-analysis with Dr. Amanda Sargent under Dr. George Banks' supervision.
Chris N. H. Street	0	0	1	0	0	0	0	0	1	0	0	0	0	1	I put together R scripts for conducting analyses and interpreted the results in light of hypotheses. I intend to contribute to the drafting once it is made available.
Eirik Strømland	0	0	0	0	0	0	0	0	0	0	1	0	0	1	Served as a reviewer on many preregistration drafts for direct replication studies, audited many of the final reports checking for errors, and reviewed and edited the final report.
Tina Sundelin	0	0	0	0	0	0	0	0	0	0	1	0	0	1	I reviewed pre-registrations for five replications, and reviewed and edited the final report.
Moin Syed	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I designed and executed one replication project, Wetzel et al. (2016), European Journal of Personality
Anna Szabelska	0	0	1	0	1	1	0	0	1	0	1	0	0	0	
Barnabas Szaszi	0	0	0	0	1	0	0	0	0	0	0	0	0	1	We conducted a replication research for Mason et al. 2012
Ewa Szumowska	0	0	1	0	1	1	0	0	1	0	0	0	0	0	Together with Aneta Czernatowicz-Kukuczka we conducted a replication of an on-site experiment by Lee et al. It included preparation of research materials, coding the procedures, data collection, recruitment of RAs, and data analysis.

Anirudh Tagat	0	0	1	0	1	1	0	0	1	0	0	0	0	0	Abouk, R., & Heydari, B. (2021). The immediate effect of COVID-19 policies on social-distancing behavior in the United States. Public Health Reports, 136(2), 245-252. [OSF Project], joint with Varsha Ashok (Royal Holloway). Anderson, S. (2011). Caste as an Impediment to Trade. American Economic Journal: Applied Economics, 3(1), 239-63. [OSF Project], joint with Nathaniel Porter and Jing Geng (Virginia Tech) Gerhold, L. (2020, March 25). COVID-19: Risk perception and Coping strategies. https://doi.org/10.31234/osf.io/xmpk4 [OSF Project], joint with Hansika Kapoor (Monk Prayogshala) Thames, F. C., & Williams, M. S. (2010). Incentives for personal votes and women's representation in legislatures. Comparative Political Studies, 43(12), 1575-1600. [OSF Project], joint with Arathy Puthillam (Monk Prayogshala)
Susanne Täuber	0	0	1	0	1	1	0	0	0	0	0	0	1	1	Our team of two people performed an experimental study with the aim to replicate a Research Claim from Bhattacharjee, Dana, & Baron (2017), from The Journal of Personality and Social Psychology. We designed and programmed the experiment, collected and analyzed the data, and wrote a report. I carried out a replication study to its completion together with my collaborator Dr. Laetitia Mulder (University of Groningen, The Netherlands).
Louis Tay	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I supervised and edited 1 replication submission for "Gabriel et al. (2011), from the Journal of Applied Psychology" working with my two former graduate students Louis Hickman and Stuti Thapa.
Stuti Thapa	0	0	1	0	0	0	0	0	0	0	0	0	0	0	I worked with Dr. Louis Tay and Dr. Louis Hickman on the Gabriel et al. (2011) replication project. I conducted all of the analyses and worked with the statistical team to finalize the models and analytical needs for the study.
Jason Thatcher	0	0	0	0	1	0	0	0	0	0	0	0	0	0	Helped to gather data
Domna Tsaklakidou	1	0	0	0	0	1	0	0	0	1	0	0	0	1	2 replication edited
Lars Tummers	0	0	1	0	1	0	0	0	0	0	1	0	0	0	Together with dr. Robin Bouwman, I carried out the replication McCarter_OrgBehavior_2010_plLK. I reviewed Fritz_JournOrgBehavior_2010_zekm.
Elise Turkovich	0	0	1	0	1	0	0	0	0	0	0	0	0	0	Helped operationalize the replication of the original study on time perception and the visual cortex in a new lab. Collected data from human subjects, often alongside Soorya Sankaran. helped organize and analyze data alongside Audrey Morrow, and Jason Samaha.
Melba Verra Tutor	0	0	0	0	1	0	0	0	0	0	0	0	0	0	I was a data finder for a number of replication studies, but I'm afraid I am not aware which ones moved forward to the analysis stage.
Karolina Urbanska	0	0	0	0	1	1	0	0	0	0	1	0	0	0	Led multiple projects - reviewing, finding datasets, preparing prereg, analysing data, reporting. Also involved in identifying claims in the earlier stage before replication kicked-off. Helped with auditing the results at the end as well.
Anna Elisabeth van 't Veer	0	0	0	0	0	0	0	0	0	0	1	0	0	0	I served as a reviewer for replication designs, but I'm sorry I cannot remember exactly anymore
Marcel van Assen	0	0	1	0	0	1	0	0	0	0	0	0	0	0	Involved in power-analyses and computation of effect sizes.
Niels van de Ven	0	0	1	0	1	1	0	0	0	0	1	0	0	1	Worked on 1 direct replication (z106). Designed study, got IRB approval, collected data. Together with Constant Pieters analyzed data and wrote replication report.
Ruben van den Goorbergh	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
Elisabeth Julie Vargo	0	0	1	0	1	1	0	0	1	0	1	0	1	1	Replication of a research claim from Van Gastel et al. (2013) in Psychological Medicine; Replication of two research claims from Button & Worthen (2017) in Criminology; Generalizability study and Replication of a research claim (Study 2) from Torelli & Shavitt (2010); Replication of a research claim (Study 3) from Usta & Häubl (2011), in Journal of Marketing Research. I reviewed probably about a dozen replication studies, if not more (I didn't record and would take me a bit of sifting through emails to provide a more specific account of the reviews).

Leigh Ann Vaughn	0	0	1	0	1	1	0	0	1	0	0	0	0	0	Kathryn Caldwell and I developed the replication of the research claim from Ku and Zaroff (2014), from the Journal of Environmental Psychology. Our Ithaca College subcontract number was Ithaca_4zy8. We designed the replication study, developed and revised the preregistration, appear to have been the first to go through ethics review for a SCORE replication (including an initial Department of the Navy (DON) Human Research Protection Office (HRPO) in fall 2019), collected and analyzed data, and submitted and revised our final report to SCORE.
Simine Vazire	0	0	0	0	0	1	0	0	0	0	0	0	0	0	Editor of preregistrations/plans for replication studies
Jentien M. Vermeulen	0	0	0	0	1	0	0	0	0	0	0	0	0	1	I have prepared data and collaborated in designing the investigation for 1 replication study
Diem Thi Hong Vo	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I was part of a team at the University of Canterbury who were involved in replications/reproductions of approximately 40 studies. We were involved in multiple aspects of the project, including data preparation, data analysis, and writing up of final reports for each study.
Victor Volkman	0	0	1	0	1	0	0	0	1	0	1	0	1	1	I took part in a replication of Liang, Lazear, and Wang (2016) and Benjamin, Choi, and Strickland (2010). In the former, I performed data analysis on additional data gathered regarding Entrepreneurship data in the surveyed countries in the years following the initial paper. I took a much more active role in the latter, not only helping adapt the questions from the original experiment to fit a segment of the general population, but programming the Qualtrics survey that allowed the experiment to be conducted online, helping organize meetings with the survey firm responsible for its implementation, and conducting the data analysis after the results were gathered.
Eric-Jan Wagenmakers	0	0	0	0	0	0	0	0	0	1	1	0	0	0	I have helped edit six replication submissions and I have had an advisory role for one replication attempt.
Deliah Wagner	1	0	1	0	1	1	0	0	1	0	1	1	1	1	I participated in the Replication of a Research Claim from Ihme & Tausendpfund (2018) from The Journal of Experimental Political Science. COS code: Ihme_JournExpPoliSci_2018_xYbO_yk16. Which was published at the same journal: DOI: https://doi.org/10.1017/XPS.2022.35
Lukasz Walasek	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I was part of a team of 4 researchers who conducted a replication of one empirical result to be included in the SCORE project. Project ID: de Langhe_JournConsRes_2014_RJb - Walasek_6m4k
Frank Walter	0	0	0	0	1	0	0	0	0	0	0	0	0	0	Together with Roman Briker, I replicated Study 1a of Reh et al. (2018, JAP). Both authors designed the replication study together. Roman Briker prepared the OSF archive for the study, performed data collection, analysis, and drafted the first version of the manuscript summarizing the replicaton (Briker & Walter, 2021, Social Psychology). Frank Walter provided feedback throughout data collection and analysis, and critically edited and revised the resulting manuscript.
Lara Warmelink	0	0	0	0	1	0	0	0	0	0	0	0	0	0	I worked as a data finder on the Boehm replication. I found the data and wrote R-code to make the data ready for analysis by another team.
Liuqing Wei	0	0	0	0	1	0	0	0	0	0	0	0	0	0	I collaborated with Professor Thomas Talhlem in the University of Chicago and helped collect data for the Ma-Kellams study replication
Marie Isabelle Weißflog	0	0	0	0	1	0	0	0	1	0	0	0	0	0	I was involved in programming the survey setup for 1 replication study, and in participant recruitment and management in the same study
Nicholas Weller	0	0	1	0	1	0	0	0	0	0	0	0	1	1	Michelangelo Landgrave and I worked together to replicate Einstein and Glick's (2017) study about discrimination by bureaucrats working for public housing agencies in the United States. We wrote the replication plan, conducted the experiment, analyzed the data, and then wrote up the results afterwards.
Aaron L. Wichman	0	0	1	0	1	1	0	0	1	1	1	1	0	1	It's really kind of a blur to me. I know I conducted the replication of Steinmetz_JournPerSocPsy_2016_E4Am_m5y9, and a replication and robustness check on Luttrell et al. 2016. I try to say yes to whatever opportunities from SCORE come along.
Jonathan Wilbiks	0	0	1	0	1	1	0	0	1	0	1	0	0	0	I conducted 3 replication projects, including creation of program, data collection, and analysis of data. I served as reviewer for 6 replication submissions.
Jamal R. Williams	0	0	1	0	1	1	0	0	0	0	0	0	0	0	We conducted a reproduction of the research claim(s) in "Asymmetrical Body Perception: A Possible Role for Neural Body Representations", by Linkenauger, et al. (2009)
Kelly Wolfe	0	0	1	0	1	1	0	0	0	0	0	0	0	0	I worked together with Miroslav Sirota and Marie Juanchich on a sure vs risky choice replication study. I created the study materials, wrote part of the code for analysis, and analysed the data, as well as trouble shooted when there were issues with the data.
Finnian Wort	0	0	0	0	0	1	0	0	0	0	0	0	0	0	I built the surveys which we used to gather data

Ryan Wright	0	0	1	0	0	0	0	0	0	0	0	0	0	0	I participated in data collection for the Johnson et al. (2012) replication.
Jesper N. Wulff	0	0	1	0	0	0	0	0	1	0	0	0	0	0	I served as a data analyst for the replication of a research claim from Subramanian (2013) from Management Science. The replication team consisted of myself and Anna Abatayo. Action Editor was Miguel Fonseca and Sorin Valcea was an independent reviewer. I received instructions from Andrew Tyner and I also had the pleasure of coordinating with the SCORE coordinator Zach and Melissa Kline.
Xindong Xue	0	0	0	0	0	0	0	0	0	0	1	0	0	0	Data Preparation work for the Fox (JOLE) paper using the SIPP Data.
Veronica X. Yan	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I was of part one of the replication studies (Muis_ContEduPsych_2009_dqKX - Schuetze - Direct Replication - 7965)
Yuzhi Yang	0	0	0	0	1	0	0	0	0	0	0	0	0	0	I assisted with data collection under the supervision of Dr. Johnathan Wilbiks.
Sangsuk Yoon	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I participated in two replication projects. I participated in the project as a replication team for two replication projects: LeBoeuf_JournMarketRes_2010_EKBZ and Zhang_ExpEco_2009_1VeW.
Iris Žeželj	0	0	1	0	1	1	0	0	0	0	0	0	0	0	I collected and analyzed data for one replication study.
Yinxian Zhang	0	0	1	0	1	0	0	0	1	0	0	0	0	0	I found secondary data, planned, and conducted a replication study of Weidmann and Callen (2013) (SCORE RR ID: 2g7ky). The study evaluates 4 claims from the original article. Upon its completion, I wrote a final report summarizing the replication process and findings.
Ignazio Ziano	0	0	0	0	0	1	0	0	0	0	1	0	0	0	I served as a reviewer for 5 replication preregistrations (including Wentzel et al. and Savani et al., but I cannot find the other ones) and 2 reproduction attempts.
Cristina Zogmaister	0	0	0	0	0	0	0	0	0	1	1	0	0	0	I servedas Editor an for one paper that contained a replication and a reproduction, and I also was a Reviewer for 42 replication studies (plus one that contained a replication and a reproduction).
Zorana Zupan	0	0	0	0	0	0	0	0	0	0	1	0	0	0	I have served as a reviewer for 5 replication submissions, and 3 reproduction submissions. (Zhou et al., 2014, Morewedge et al., 2009, Smith et al., 2016, Muis et al., 2009, Seaton et al., 2010, Roberts et al, 2010, Al Tammemi et al, 2020, Travers&Kreizman, 2018)
Rolf A. Zwaan	0	0	0	0	0	0	0	0	0	0	1	0	0	0	I served as a reviewer on a number of submissions but cannot remember how many.
Brian A. Nosek	1	0	0	1	0	1	1	0	0	1	0	1	1	1	PI of the TA1 team from the SCORE program (Center for Open Science). Contributed high-level design, visioning, and leadership for the project. Collaborated closely with the COS project leader (Tim Errington) on COS's contribution to the program. Coordinated across teams on project planning, executing, and reporting.
Timothy M. Errington	1	0	0	1	1	1	1	0	0	1	1	0	1	1	Project lead of the TA1 team from the SCORE program (Center for Open Science). Contributed to high-level design, visioning, leadership, and operationalization for the project. Coordinated across teams on project planning, executing, and reporting.

References

1. Nosek, B. A. & Errington, T. M. What is replication? *PLOS Biol.* **18**, e3000691 (2020).

2. National Academies of Sciences, E. *Reproducibility and Replicability in Science*. (2019). doi:10.17226/25303.

3. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **349**, aac4716–aac4716 (2015).

4. Klein, R. A. *et al.* Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Adv. Methods Pract. Psychol. Sci.* **1**, 443–490 (2018).

5. Klein, R. A. *et al.* Investigating Variation in Replicability: A “Many Labs” Replication Project. *Soc. Psychol.* **45**, 142–152 (2014).

6. Ebersole, C. R. *et al. Many Labs 5: Testing Pre-Data Collection Peer Review as an Intervention to Increase Replicability (Results-Blind Manuscript)*. (2019). doi:10.31234/osf.io/sxfm2.

7. Ebersole, C. R. *et al.* Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *J. Exp. Soc. Psychol.* **67**, 68–82 (2016).

8. Camerer, C. F. *et al.* Evaluating replicability of laboratory experiments in economics. *Science* **351**, 1433–1436 (2016).

9. Camerer, C. F. *et al.* Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat. Hum. Behav.* **2**, 637–644 (2018).

10. Nosek, B. A. *et al.* Replicability, Robustness, and Reproducibility in Psychological Science. *Annu. Rev. Psychol.* (2021).

11. Cova, F. *et al.* Estimating the Reproducibility of Experimental Philosophy. *Rev. Philos. Psychol.* **12**, (2018).

12. Errington, T. M. *et al.* Investigating the replicability of preclinical cancer biology. *eLife* **10**, e71601 (2021).

13. Munafò, M. R. *et al.* A manifesto for reproducible science. *Nat. Hum. Behav.* **1**, 0021 (2017).

14. Ioannidis, J. P. A. Why Most Published Research Findings Are False. *PLoS Med.* **2**, e124 (2005).

15. Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).

16. Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. & Kievit, R. A. An agenda for purely confirmatory research. *Perspect. Psychol. Sci.* **7**, 632–638 (2012).

17. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychol. Sci.* **22**, 1359–1366 (2011).

18. Greenwald, A. G. Consequences of prejudice against the null hypothesis. *Psychol. Bull.* **82**, 1–20 (1975).

19. Rosenthal, R. The file drawer problem and tolerance for null results. *Psychol. Bull.* **86**, 638–641 (1979).

20. Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. The preregistration revolution. *Proc. Natl. Acad. Sci.* **115**, 2600–2606 (2018).

21. Giner-Sorolla, R. Science or Art? How Aesthetic Standards Grease the Way Through the Publication Bottleneck but Undermine Science. *Perspect. Psychol. Sci.* **7**, 562–571 (2012).

22. Nosek, B. A., Spies, J. R. & Motyl, M. Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspect. Psychol. Sci.* **7**, 615–631 (2012).

23. Nosek, B. A. *et al.* Promoting an open research culture. *Science* **348**, 1422–1425 (2015).

24. Freese, J. & Peterson, D. Replication in Social Science. *Annu. Rev. Sociol.* **43**, 147–165 (2017).

25. Andrews, I. & Kasy, M. Identification of and correction for publication bias. *Am. Econ. Rev.* **109**, 2766–94 (2019).

26. Patil, P., Peng, R. D. & Leek, J. T. What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspect. Psychol. Sci.* **11**, 539–544 (2016).

27. Valentine, J. C. *et al.* Replication in Prevention Science. *Prev. Sci.* **12**, 103–117 (2011).

28. Abatayo, A. L. *et al.* Empirical, Human, and Machine Assessments of Research Credibility in the Social and Behavioral Sciences. *Preprint* (2025).

29. Fanelli, D. “Positive” Results Increase Down the Hierarchy of the Sciences. *PLoS ONE* **5**, e10068 (2010).

30. Fanelli, D. Negative results are disappearing from most disciplines and countries. *Scientometrics* **90**, 891–904 (2012).

31. Heyard, R. *et al.* A scoping review on metrics to quantify reproducibility: a multitude of questions leads to a multitude of metrics. *MetaArXiv* (2024).

32. Muradchanian, J., Hoekstra, R., Kiers, H. & van Ravenzwaaij, D. How best to quantify replication success? A simulation study on the comparison of replication success metrics. *R. Soc. Open Sci.* **8**, 201697 (2021).

33. Nosek, B. A. *et al.* Replicability, Robustness, and Reproducibility in Psychological Science. *Annu. Rev. Psychol.* **73**, 719–748 (2022).

34. Peels, R. Replicability and replication in the humanities. *Res. Integr. Peer Rev.* **4**, 2 (2019).

35. Peels, R. & Bouter, L. The possibility and desirability of replication in the humanities. *Palgrave Commun.* **4**, 95 (2018).

36. TalkadSukumar, P. & Metoyer, R. *Replication and Transparency of Qualitative Research from a Constructivist Perspective*. https://osf.io/6efvp (2019) doi:10.31219/osf.io/6efvp.

37. Hubbard, D. W. & Carriquiry, A. L. Quality Control for Scientific Research: Addressing Reproducibility, Responsiveness, and Relevance. *Am. Stat.* **73**, 46–55 (2019).

38. Alipourfard, N. *et al.* Systematizing Confidence in Open Research and Evidence (SCORE). (2021).

39. Miske, O. *et al.* Investigating the reproducibility of the social and behavioral sciences. *Preprint* (2025).

40. Abatayo, A. L. *et al.* Overview of the SCORE Program Methodology and Reporting. *Preprint* (2025).

41. Ebersole, C. R. *et al.* Many Labs 5: Testing Pre-Data-Collection Peer Review as an Intervention to Increase Replicability. *Adv. Methods Pract. Psychol. Sci.* **3**, 309–331 (2020).

42. Held, L., Pawel, S. & Micheloud, C. The assessment of replicability using the sum of p-values. *R. Soc. Open Sci.* **11**, 240149 (2024).

43. Micheloud, C., Balabdaoui, F. & Held, L. Assessing replicability with the sceptical p-value: Type-I error control and sample size planning. *Stat. Neerlandica* **n/a**,.

44. Gronau, Q. F., Heck, D. W., Berkhout, S. W., Haaf, J. M. & Wagenmakers, E.-J. A Primer on Bayesian Model-Averaged Meta-Analysis. *Adv. Methods Pract. Psychol. Sci.* **4**, 25152459211031256 (2021).

45. Simonsohn, U. Small Telescopes: Detectability and the Evaluation of Replication Results. *Psychol. Sci.* **26**, 559–569 (2015).

46. Verhagen, J. & Wagenmakers, E.-J. Bayesian tests to quantify the result of a replication attempt. *J. Exp. Psychol. Gen.* **143**, 1457–1475 (2014).

47. Ly, A., Etz, A., Marsman, M. & Wagenmakers, E.-J. Replication Bayes factors from evidence updating. *Behav. Res. Methods* **51**, 2498–2508 (2019).

48. Steiner, P. M., Sheehan, P. & Wong, V. C. Correspondence measures for assessing replication success. *Psychol. Methods* No Pagination Specified-No Pagination Specified (2023) doi:10.1037/met0000597.

49. Ben-Shachar, M., Lüdecke, D. & Makowski, D. effectsize: Estimation of Effect Size Indices and Standardized Parameters. *J. Open Source Softw.* **5**, 2815 (2020).

Supporting Information for "Investigating the replicability of the social and behavioral sciences"

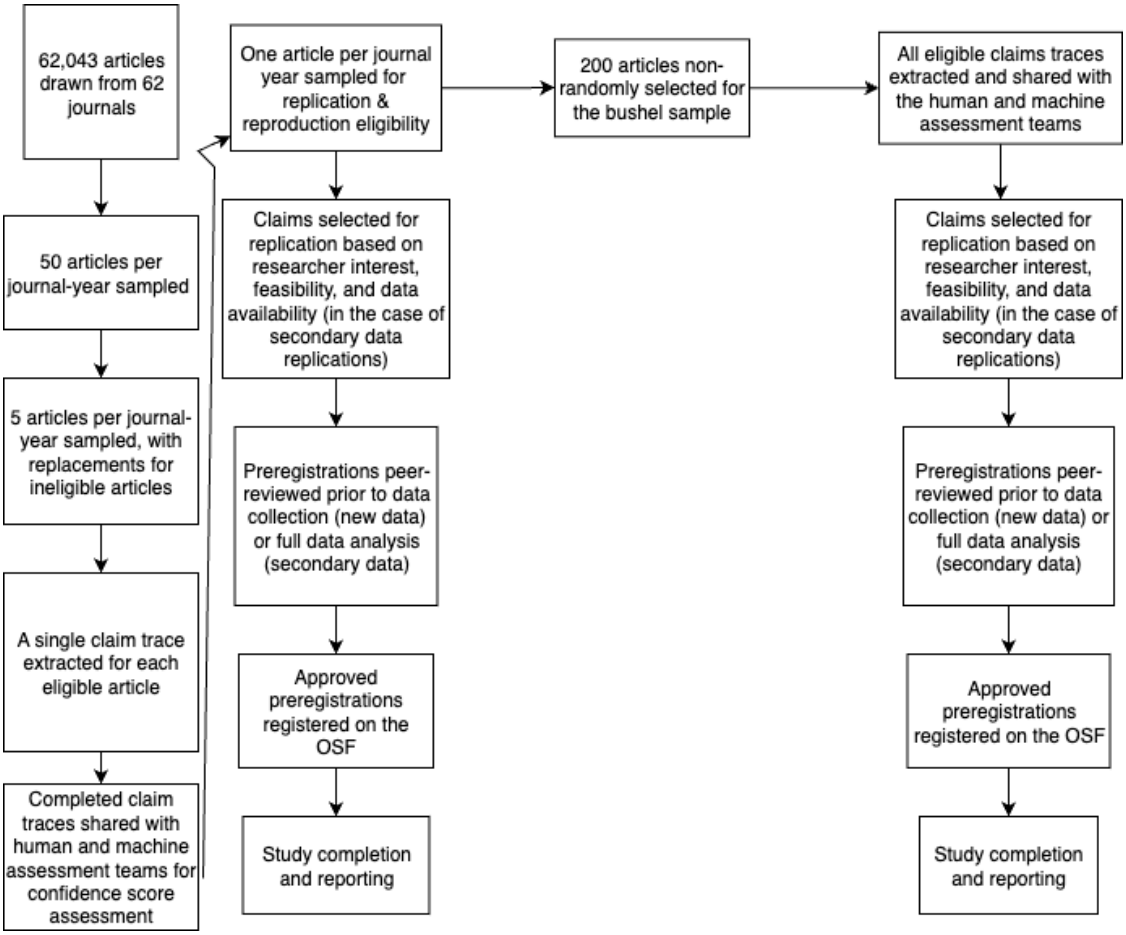
Andrew H. Tyner, Anna Lou Abatayo, Mason Daley, Samuel Field, Nicholas Fox, Noah A. Haber, Krystal M. Hahn, Melissa Kline Struhl, Brinna Mawhinney, Olivia Miske, Priya Silverstein, Courtney K. Soderberg, Theresa Stankov, Ahmed Abbasi, Christopher L. Aberson, Balazs Aczel, Matúš Adamkovič, Nihan Albayrak, Peter J Allen, Michael Andreychik, Eli Awtrey, Erick Axxe, Flavio Azevedo, Miles D. Bader, Bence Bago, James Bailey, Marjan Bakker, Gabriel Banik, George C. Banks, Ernest Baskin, Anatolia Batruch, Annika Beatteay, Sophie M. Behr, Nicholas Berente, Zachariah Berry, Jędrzej Białkowski, Bojana Bodroža, Laura Boeschoten, Miklos Bognar, Christian Bokhove, Diane Bonfiglio, Robin Bouwman, Timothy F. Brady, Scott Braithwaite, Gabriel Briceño Jiménez, Cameron Brick, Traci Bricka, Roman Briker, Annette N. Brown, Gordon D A Brown, Robbie C.M. van Aert, Kathryn Caldwell, Sara Capitan, Tabaré Capitán, Jesse Chandler, Tessa Charles, Christopher R. Chartier, Rahul Chawdhary, Kent Jason Cheng, William J. Chopik, Bruce Clark, Victoria E. Colvin, C. Cozette Comer, Giulio Costantini, Tom Coupé, Jamie Cummins, Aneta Czernatowicz-Kukuczka, Joshua de Leeuw, David Dobolyi, James N. Druckman, Jianhua Duan, Marin Dujmović, Daniel J. Dunleavy, Patrick K. Durkee, Cécile Emery, Kevin M. Esterling, Thomas R. Evans, Anna Fedor, Belén Fernández-Castilla, Nathan Fiala, James G. Field, Nathan Fong, Miguel A. Fonseca, Alexandra L.J. Freeman, Jeremy Freese, Sandra J. Geiger, Jing Geng, Laura M. Getz, Linda Marjoleine Geven, Ilka Helene Gleibs, Donna Pamella Gonzales, Janaki Gooty, Amélie Gourdon-Kanhukamwe, Cristina Greculescu, Siobhán M. Griffin, Lusine Grigoryan, Martina Grunow, Nicholas Gunby, Braeden Hall, Paul H. P. Hanel, Erin E. Hannon, Sam Harper, Marco Jürgen Held, Louis Hickman, Nathan C. Higgins, Svenja Hippel, Sven Hoeppner, Sanghyun Hong, Thomas J. Hostler, Michael Inzlicht, Kamil Izydorczak, Bastian Jaeger, Kristin Jankowsky, Johannes Jarke-Neuert, Matthew Jensen, Biljana Jokić, Daniel Jolles, Phillip Jolly, Angela M. Jones, Marie Juanchich, Pavol Kačmár, Hansika Kapoor, Andjela Keljanovic, Samjhana Koirala, Marta Kołczyńska, Dimitra Kouroupaki, Ulrich Kühnen, Michelangelo Landgrave, Michael J. Larson, Lyonel Laulié, Alice C E Lawrence, Joel M. Le Forestier, Katelin E. Leahy, Sungmok Lee, Jared Leslie, Savannah C. Lewis, Christopher Limnios, Hause Lin, An-Chiao Liu, John Wills Lloyd, Elliot A Ludvig, Dermot Lynott, Jordan MacDonald, Peter Mallik, Daniel J. Mallinson, Daniele Marinazzo, Corinna S. Martarelli, Joshua Matacotta, Andrew McBride, Cillian McHugh, Gail McMillan, Esteban Méndez, Mitchell Metzger, Michalis P. Michaelides, Johannes Michalak, Leticia Micheli, Jeremy K. Miller, Marina Milyavskaya, Daniel C. Molden, Ambar G. Monjaras, David Moreau, Audrey Morrow, Cristóbal Moya, Liad Mudrik, Laetitia B. Mulder, Katie A. Munt, Arijit Nandi, Kathryn Nason, Carolin Nast, Gideon Nave, Heinrich H. Nax, Florian Neubauer, Phuong Linh L. Nguyen, Austin Lee Nichols, Gustav Nilsonne, Ernest O'Boyle, Jule Oettinghaus, Jeewon Oh, Adoril Oshana, Thomas Ostermann, Rachel P. Ostrowski, Abiola Oyebanjo, Radoslaw Panczak, Jamie Patrianakos, Ignacio Pavez, Yuri G. Pavlov, Sofia Persson, Marco Perugini, Kim Peters, Constant Pieters, Vladimir Ponizovskiy, Nathaniel D. Porter, Jason M. Prenoveau, Danka Purić, Mariah F. Purol, Arathy Puthillam, Kimberly A. Quinn, Marco Ramljak, W. Robert Reed, Michaela Ritchie, Margaret Ritzau, Sean Patrick Roche, Romina Rodela, Jan Philipp Röer, Ivan Ropovik, Jacob Rothschild, Justine Saal, Hani Safadi, Jason Samaha, Mary Sanchez, Soorya Sankaran, David Santos, Amanda C. Sargent, Marian Sauter, Kathleen Schmidt, Landon Schnabel, Amber N Schroeder, Sebastian W. Schuetz, Brendan A. Schuetze, Michael Schulte-Mecklenbeck, Astrid Schütz, Eric L. Sevigny, Ellie Shackleton, Richard M. Shafranek, Samuel Shaki, Shishir Shakya, Miroslav Sirota, Matthew Ryan Sisco, Maksim M. Sitnikov, L. Robert Slevc, Laura Smalarz, Colin Tucker Smith, Joel S. Snyder, Nicolas Sommet, Fatih Sonmez, Barbara A. Spellman, Natalia Stanulewicz-Buckley, George Stock, Chris N. H. Street, Eirik Strømland, Tina Sundelin, Moin Syed, Anna Szabelska, Barnabas Szaszi, Ewa Szumowska, Anirudh Tagat, Susanne Täuber, Louis Tay, Stuti Thapa, Jason Thatcher, Domna Tsaklakidou, Lars Tummers, Elise Turkovich, Melba Verra Tutor, Karolina Urbanska, Anna Elisabeth van 't Veer, Marcel van Assen, Niels van de Ven, Elisabeth Julie Vargo, Leigh Ann Vaughn, Simine Vazire, Jentien M. Vermeulen, Diem Thi Hong Vo, Victor Volkman, Eric-Jan Wagenmakers, Deliah Wagner, Lukasz Walasek, Frank Walter, Lara Warmelink, Liuqing Wei, Marie Isabelle Weißflog, Nicholas Weller, Aaron L. Wichman, Jonathan Wilbiks, Jamal R. Williams, Kelly Wolfe, Finnian Wort, Ryan Wright, Jesper N. Wulff, Xindong Xue, Veronica X. Yan, Yuzhi Yang, Sangsuk Yoon, Iris Žeželj, Yinxian Zhang, Ignazio Ziano, Cristina Zogmaister, Zorana Zupan, Rolf A. Zwaan, Brian A. Nosek, & Timothy M. Errington

Table of Contents

Overview

This supplement provides details on the methodology and additional results for the replications attempted during the SCORE program funded by DARPA. Replication studies were just one component of the SCORE program. Background on the design and approach of the whole program is available in (Abatayo et al., 2025). Replications were conducted on claims extracted from a sample of papers published in social and behavioral science journals. Details about the identification and extraction of claims is available in (Abatayo et al., 2025). A visual overview of the sampling and replication assessment process is in Figure S1.

Figure S1. Workflow diagram of sampling and replication process. Details of the methods are reported in this paper and supporting information and in Abatayo et al. (2025).



Methods

The "Systematizing Confidence in Open Research and Evidence" (SCORE) program relied on a dataset of replication outcomes to serve as ground truth evidence to validate human and AI predictions about the credibility of social and behavioral science claims. This dataset was produced in a coordinated, distributed effort of researchers across the globe. Individual researchers and small teams provided their substantive expertise to conduct high-quality, good faith replication attempts, and a coordinating team provided the operational, financial, and logistical support to maintain the timeline; facilitate an internal peer review process to ensure

projects were rigorously conducted; provide just-in-time statistical consulting for power analyses; and conduct training in best practices for data sharing, preregistration, and outcome reporting for consistency across the project.

The priorities of this replication effort were rigor, transparency, and efficiency. Accordingly, this document focuses on those aspects of the process which facilitated these goals and gives relatively less attention to specific operational steps which are less relevant to understanding SCORE's replication evidence such as grant making and managing the cooperative agreement with DARPA. The procedures documented below evolved over the three years of SCORE. When relevant to interpreting SCORE evidence, we highlight changes to our processes and indicate when in the project's timeline the changes were implemented.

Several parts of the methods for this supporting paper are very similar to the supporting information for SCORE reproductions because of a shared methodology (Miske et al., 2025). For some sections, the same initial description was used and then edited separately for the features unique to the replication and reproduction attempt methods.

Key Terms

Abatayo et al. (2025) provide a glossary of key terms for the program. Here, we highlight a few that are particularly important for understanding the replication methodology. Replication attempts involve testing the same claim as an original paper with different data. Replication attempts could be conducted on independent secondary data or on new data collected during the program. For purposes idiosyncratic to the program, new data replications were labeled "direct replications" or "DRs" in some of the historical documentation and secondary data replications were labeled "data-analytic replications" or "DARs". For reporting purposes, we use the terms "new data replications" and "secondary data replications" but the alternate terms might appear in older documentation.

There was a third category of replication evidence that included a combination of original observations and independent observations not used in the original analysis. These are called "hybrid replications." Because they were not based on completely independent evidence, hybrid replication attempts were not included in the dataset reported in the main text. For the reader's interest, they are reported separately in this supporting material.

An idiosyncratic feature of the program was that the research was conducted in two phases. This is a common feature of DARPA programs in which progress in one phase is used to determine whether to fund a next phase. We report cumulative evidence gathered across the phases. On occasion, procedures were updated between Phase 1 and Phase 2. Those changes are highlighted if they could have substantive implications for the methods, results, or findings.

Sample and Data

Claims for potential replication were drawn from 27407 social and behavioral science papers published from 2009 to 2018. Table S1 presents the 62 journals included in the sampling frame for selecting papers and claims. Additional details about the journal selection process, paper selection process, and claim extraction process are reported in Abatayo et al. (2025). Here, we provide information to supplement the main text for understanding selection of papers and claims for replication attempts.

The project was conducted in two phases. Most of the replication attempts came from papers selected in Phase 1 (3000 vs 900 in Phase 2) as were most of the completed replications (139 vs 25 in Phase 2). Also, we reduced the Phase 1 sample of papers eligible for attempting

replications with a second round of random sampling: We reduced 3000 to 600 papers to evaluate selection effects more effectively. Of that 600, 200 had multiple claims and 400 had just a single claim extracted. Given program time constraints in Phase 2, we intentionally attempted to replicate only 25 of the 900 eligible papers, all with a single claim extracted. Programmatically, we focused our efforts on deepening the evidence collected for Phase 1 papers during Phase 2. In the Results section below, we reproduced summary findings from the main text separately for Phase 1 and Phase 2 data.

During the program, a subset of 200 papers from the sample of 600 identified in Phase 1 was created non-randomly, focusing first on papers for which replication or reproduction evidence had been gathered, second on papers likely to be able to produce replication and reproduction evidence, and third on retaining relative representatives of papers across disciplines in the subset of 200 papers. The 200 papers went through an additional claim extraction process in Phase 2 to code all eligible claims from those papers instead of just a single claim.

Table S1. 62 journals included in the sampling frame for selecting papers and claims.

Business	**Education**	**Psychology**
Academy of Management Journal	American Educational Research Journal	Child Development
Journal of Business Research	Computers and Education	Clinical Psychological Science
Journal of the Academy of Marketing Science	Contemporary Educational Psychology	Cognition
Journal of Consumer Research	Educational Researcher	European Journal of Personality
Journal of Marketing	Exceptional Children	Evolution and Human Behavior
Journal of Marketing Research	Journal of Educational Psychology	Health Psychology
Journal of Management	Learning and Instruction	Journal of Applied Psychology
Journal of Organizational Behavior		Journal of Consulting and Clinical Psych.
Leadership Quarterly		Journal of Environmental Psychology
Management Science		Journal of Experimental Psychology: General
Organization Science		Journal of Experimental Social Psychology
Organizational Behavior and Human Decision Processes		Journal of Personality and Social Psychology
		Psychological Science
		Psychological Medicine
		Social Science and Medicine

Economics	**Political Science**	**Sociology**
American Economic Journal: Applied Economics	American Journal of Political Science	American Journal of Sociology
American Economic Review	American Political Science Review	American Sociological Review
Econometrica	British Journal of Political Science	Criminology
Experimental Economics	Comparative Political Studies	Demography
Journal of Finance	Journal of Conflict Resolution	European Sociological Review
Journal of Financial Economics	Journal of Experimental Political Science	Journal of Marriage and Family
Journal of Labor Economics	Journal of Public Administration Research and Theory	Law and Human Behavior
Journal of Political Economy	Public Administration Review	Social Forces
Quarterly Journal of Economics	World Politics	
Review of Financial Studies		
World Development		

Caption: For primary reporting, Economics and Finance were combined as "Economics," Sociology and Criminology were combined as "Sociology," Management, Marketing, and Organizational Behavior were combined as "Business," Psychology and Health were combined as "Psychology," and Political Science and Public Administration were combined as "Political Science." Replication attempts were ultimately completed for papers from 54 of these journals.

Ethical approval

All human subjects data collection required a two-stage review to comply with both local and federal ethical requirements.

Local (IRB)

Each project received approval from the institutional review board (IRB) at the collaborating lab's home institution. We provided labs with instructions to help them prepare their project. To receive funds from the SCORE project funding agency, this institution was required to have an active Federalwide Assurance (FWA), which is an assurance that human research at that institution follows regulations outlined by the U.S. Department of Health and Human Services found in 45CFR46. If the institution of the collaborating lab did not have an active FWA, the ethical review was performed by a private IRB contracted for review called BRANY.

Federal (HRPO)

After the project received IRB approval, the project was reviewed by the Human Resource Protection Office (HRPO) of the US Army or the US Navy. We provide labs a checklist to help them prepare their project for HRPO review. While the project was under ethical review, collaborating labs prepared the research protocol. Once the HRPO approval was received, labs completed their preregistration document and submitted for internal peer review.

Sourcing Replication Teams

The recruitment of teams to attempt replications, referred to as sourcing, was facilitated by construction of a dataset of expert individuals and laboratories that represents the collective resources of the Center for Open Science (COS) through its several prior large-scale replication and reproduction projects and COS's partners: Psychological Science Accelerator (https://psysciacc.org/), the Berkeley Initiative for Transparency in the Social Sciences (BITSS; http://bitss.org/) an extensive network of economists, sociologists, political scientists, psychologists, and other social scientists, and International Initiative for Impact Evaluation (3ie; http://www.3ieimpact.org) with access to a global team of researchers from a variety of social sciences, particularly developmental economics. Each of these groups has substantial experience conducting replications or reproductions, and had expressed interest in participating in this program. Leveraging these networks meant that most researchers in the database had experience with replication or reproduction studies. Replication and reproduction attempts were sourced through the same process. The reproduction studies are reported in Miske et al. (2025).

Potential contributors to the SCORE program responded to calls for collaborators by completing a short survey about their expertise and available resources (e.g., samples, instrumentation, computing power). Survey respondents, along with individuals who were recruited through social media and word of mouth, comprised the SCORE collaborators email Google group. More than 200 researchers participated in the replication and reproduction efforts, many of whom completed multiple projects. Sourcing projects to repeat performers reduced the onboarding cost and positively contributed to the scalability of the program.

The initial intent was to manually match researchers to projects, but this approach was unnecessarily time consuming and strained coordination resources. We shifted to a more

distributed self-selection method in which performers selected projects using Google sheets that identified the topic, and highlighted unique sampling, instrumentation, setting, or expertise areas needed to conduct the research. The sheets were distributed via email to the SCORE collaborators Google group.

After signing up for a replication by providing their names and email addresses, prospective new data replication teams were instructed in a follow-up email to share their institution's FWA status and complete a Google form, confirming their eligibility to receive funding and ability to facilitate the first round of ethics review through their institution. This step ensured efficiency of the ethics review process and reduced the chances of project failure. Prospective analysts with an active FWA and local IRB were invited to an onboarding webinar and provided instructions to initiate their projects via email, beginning with the submission of an ethics review protocol and completion of a preregistration.

Secondary data replications sourced during Phase 1 involved two separate roles. A *data finder* sought relevant datasets for conducting independent replications of original claims. Data finders nominated found datasets and the coordinating team evaluated their suitability for replication. A *data analyst* then evaluated the data in detail, prepared the analysis protocol, and conducted the replication following peer review. In Phase 1, original authors received an email confirming that their paper had been matched to an analyst for conducting the replication but this was not continued in Phase 2.

In Phase 2, the process was altered slightly, as the primary emphasis shifted from new data collections to projects involving secondary data. Again, a sign-up sheet and Google form was sent to the full SCORE Collaborators group, but prospective analysts were not required to verify FWA nor IRB status. Because a majority of performers in Phase 2 had already completed projects during Phase 1, they also did not attend an onboarding webinar. In the Phase 2 sourcing form, prospective analysts were asked to investigate the feasibility of their project of interest by locating the data necessary to conduct a replication attempt. Any researcher who indicated that they could obtain access to the relevant data was sourced and provided instructions and project materials, including a link to their OSF project component, a preregistration form, and a link to the OSF component containing original materials that were either shared by the original author or located online by the coordinating team.

All projects requiring the collection of new human subjects data in Phase 2 were sourced in March and April of 2021, following email solicitation to an interested collaborator list. A total of 57 projects were tentatively sourced, meaning individuals expressed interest in attempting a project. Thirty-seven new data projects requiring human subjects review were formally sourced, meaning the collaborators started to work on the project, beginning with ethical review at their home institution and submitting their budget proposal to the coordinating team.

Power

Each replication study's target sample size was determined by conducting a power calculation prior to the study taking place. In most cases, an *a priori* power analysis was performed by a member of the coordinating team or by a statistical consultant. In a few cases, power calculations were performed by the replication team conducting the study.

Power calculations were conducted in R, using scripts developed by the coordinating team that were designed to calculate power for a variety of statistical tests and effect size types. The sample size calculations were based on information extracted directly from the original paper

(e.g., the original test statistic, effect size, p-value), however sometimes values were taken from other sources (e.g., original data or statistical code/output provided by the original author).

All effect sizes were recalculated during the power analysis, even if an effect size was reported in the original paper. This was to ensure that there wasn't an error in the original paper's reporting of an effect size (e.g., there may be ambiguity about which effect size is being reported, or conflicts between the reported test statistic and effect size), and also to ensure that we were consistent in the type of effect sizes we used across tests (e.g., there are many different types of cohen's ds, which are often all reported as simply "Cohen's d", so it was not always clear whether the original paper used the same type of effect size as we used for the power analysis).

Much of the R code uses the pwr package (v1.2-2; Champely et al., 2018) to calculate power, however other packages and approaches to estimating statistical power were used as well. The SER and SER-t method was developed to estimate power for logistic, probit, SEM, multilevel, hierarchical, and related models (see SER method description), as the traditional power analyses could not be applied in these cases.

Power calculations were performed for most replications by statistical consultants at the department of Methodology and Statistics at Tilburg University after the study was complete. These did not inform sample size definition, but did offer an opportunity to have power estimations performed consistently by an independent team. These calculations used the actual sample size in the replication attempt to estimate the power to detect the effect observed in the original study. As such, these were performed after the fact, but they are not "post hoc power" in the sense of estimating the power to detect the effect size observed in the study itself. These are the power estimates that are reported in the main text and in figures; the *a priori* power estimates are also available in shared data.

The power calculation using the SER and SER-t methods assumed that the original and replication studies had the same research design (e.g., same dependent variable including the scale with which it was measured, same correlation between dependent observations in case of a multilevel design, same number of level-1 units per level-2 unit, etc.), and only differed with respect to (total) sample size. Hence, the power was computed by adjusting the standard error of the original study for the sample size of the replication (i.e., adjusted standard error is SE_o, where SE_o is the standard error of the original study and N_o and N_r are the sample sizes of the original and replication study). The adjusted standard error was then used for calculating the power in this additional analysis.

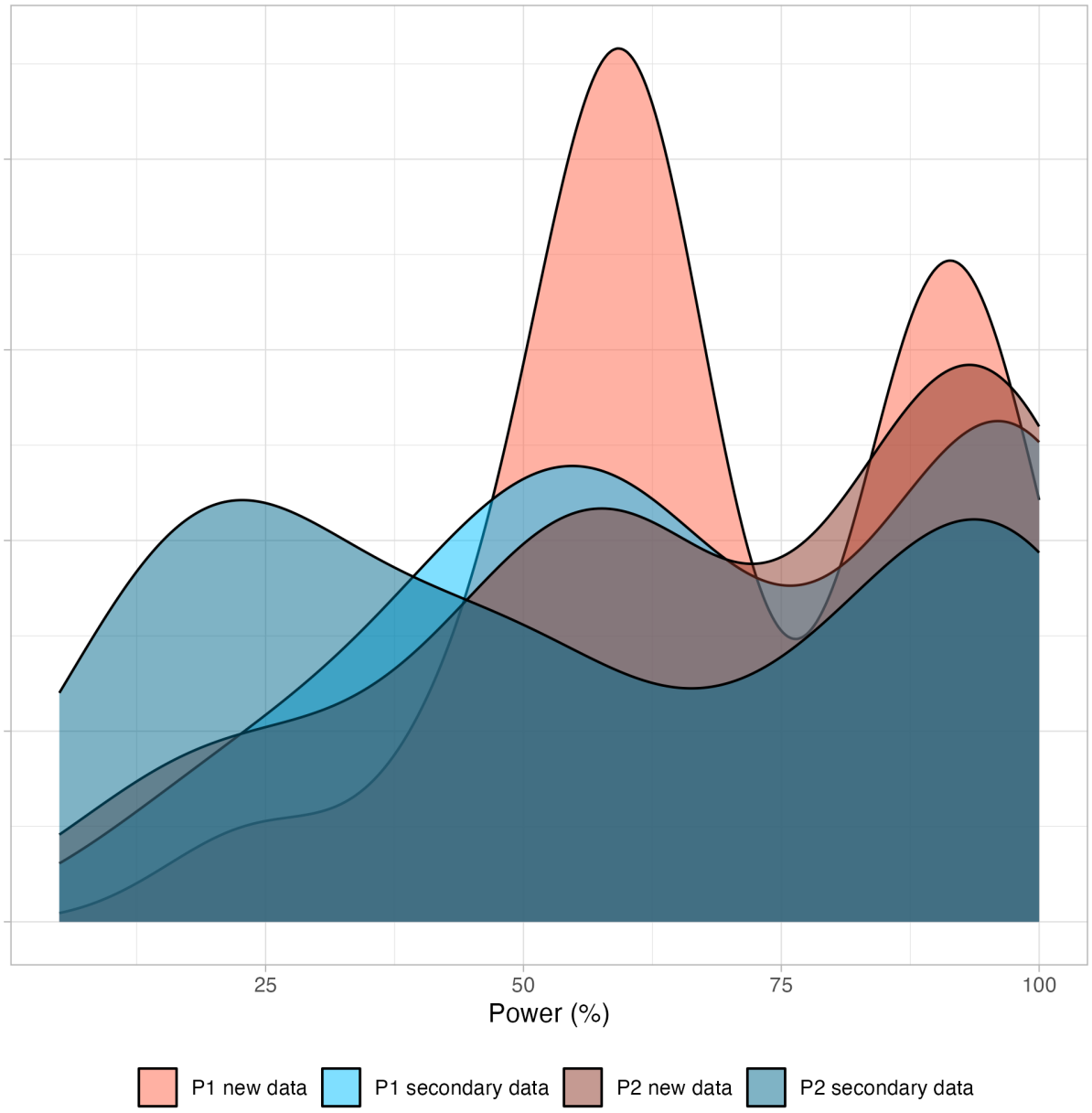
Three sample sizes were calculated for replication attempts (using an alpha level of .05 [two-tailed]; see Camerer et al. (2018) for details and impact on the false positive rate the two stage sampling process):

- A "minimum threshold sample size," defined as the sample size required to achieve 50% power to detect 100% of the original effect size.
- A "stage 1 sample size," defined as the sample size required to achieve 90% power to detect 75% of the original effect size.
- A "stage 2 sample size," defined as the sample size required to achieve 90% power to detect 50% of the original effect size.

The sampling plan and determining the target sample size varied depending on whether it was a new or secondary data replication, and whether the replication was conducted in Phase 1 or

Phase 2. Figure S2 illustrates the distribution of achieved power for different phases and types of replications described in the next sections.

Figure S2. Distribution of power to achieve 75% of the original effect size



Caption: P1 = Phase 1; P2 = Phase 2. New = new data replications; Secondary = secondary data replications.

R scripts and documentation used for conducting power calculations and documenting their outcomes and variables of interest are available in the power section in the documents folder of the OSF project. Each replication's power calculation and sample size target was documented in that study's respective preregistration and OSF project.

Sample Size Planning

Replication attempts with multiple claims

Several replication attempts were made during Phase 2 on papers that were selected during Phase 1 and in the sample of 200 for which all key claims were coded from the paper. For these cases, there could be more than one claim that would be replicated in a single replication study. We conducted a single *a priori* power analysis for a key claim and inferential test for sample size planning. We defaulted to using the single claim extracted for Phase 1 of the project if it was relevant for the replication study. Replication teams were encouraged but not required to conduct and document power calculations on the other claims and inferential tests included in their replication study. We also then conducted power calculations using the achieved combined sample size.

Estimating power for unusual cases

Nine of the new data replication attempts were recognized at the outset as using sufficiently distinct methods that the original design and effect size might not be an appropriate guide for power estimation. As such, these replication attempts could be categorized as “generalizability studies” and required an independent *a priori* power analysis that depended on the lab’s study design. As a consequence, the replication team conducted their own power analysis. We provided guidelines for determining the effect size of interest (researchers were free to use the effect size from the original study, other similar studies or relevant effect sizes from the literature, and their own judgment), conducting the power analysis (by providing them with standard power analysis scripts and encouraged their use if relevant), but set required power and alpha levels (i.e., 90% power, 5% alpha). It was ultimately up to the replication lab’s discretion to determine the effect size of interest and appropriate power calculation for their study design, and it was their responsibility to conduct the analysis with documentation and rationale in the preregistration and OSF project. These studies went through the replication peer review process just like other new data replications, and we asked reviewers to review the power analysis during this review.

New data replications conducted during Phase 1

Power calculations were done in accordance with the guidelines of the Social Sciences Replication Project (SSRP; (Camerer et al., 2018)), which defined two stages for determining sample size. In SSRP, replications that failed to achieve the original result with the Stage 1 sample size (90% power to detect 75% of the original effect size) collected additional data to achieve the Stage 2 sample size (90% power to detect 50% of the original effect size). Across all claims, 79.3% of new data replication attempts started during Phase 1 achieved 90% power to detect 75% of the original effect size, with a mean achieved power of 92.6% (median = 93.8%).

Secondary data replications conducted during Phase 1

Secondary data replications were planned to meet the “minimum threshold” sample size (i.e., 50% power to detect 100% of the original effect). Across all claims, 97.4% of secondary data replications started during Phase 1 achieved 50% power to detect 100% of the original effect size, with an average achieved power of 93.5% (median = 99.5%).

New data replications conducted during Phase 2

Power calculations were conducted in accordance with the guidelines of the Social Sciences Replication Project (SSRP; Camerer et al., 2018)) However, in a change from Phase 1, for

projects started in Phase 2, we aimed to achieve the Stage 2 sample size (90% power to detect 50% of the original effect size) from the outset. And, if this sample size was not attainable for the replicating lab, then they aimed to achieve the Stage 1 sample size (90% power to detect 75% of the original effect size). This removed the need for an interim analysis and decision about additional data collection, and encouraged replication teams to collect as much data as possible. For papers with multiple claims, only a single key claim was the basis for the power analysis. For claims with a power estimate, 72.4% of new data replication attempts started during Phase 2 achieved 90% power to detect 75% of the original effect size, with a mean achieved power of 88.1% (median = 94.1%).

Secondary data replications conducted during Phase 2

Secondary data replications were required to meet the "minimum threshold" for power that was established in Phase 1 (50% power to detect 100% of the original effect), and participating teams were encouraged to obtain the highest power possible given what data were available. We adopted this more liberal inclusion criterion for secondary data replications because secondary data replications are constrained by however much data are already available. In Phase 2, it was more common for the replication team to ensure that their attempt would meet the minimum power threshold, but each replication's power analysis or power assessment and rationale was reviewed by external researchers during the preregistration review. For papers with multiple claims, only a single key claim was the basis for the power analysis. Across all claims, 86.6% of secondary data replications started during Phase 2 achieved 50% power to detect 100% of the original effect size, with an average achieved power of 83.6% (median = 97.9%).

Combining the power approaches

The foregoing power estimates rely on a combination of the power approaches discussed earlier, which we have labeled as the traditional approach versus the SER approach. Under this combined measure, any replication outcome with a non-missing SER power estimate is assigned that value, with all other findings drawing on the traditional power estimate. In the figures below, we contrast power estimates for the two approaches across effect sizes that are 50%, 75%, and 100% of the original effect size for secondary and new data replications separately (Figures S3 and S4, respectively). We also include the distributions of power estimates across the three measures (including the combined one) for the same effect size thresholds (Figure S5).

Figure S3. Secondary data replication power estimates for SER versus traditional approach.

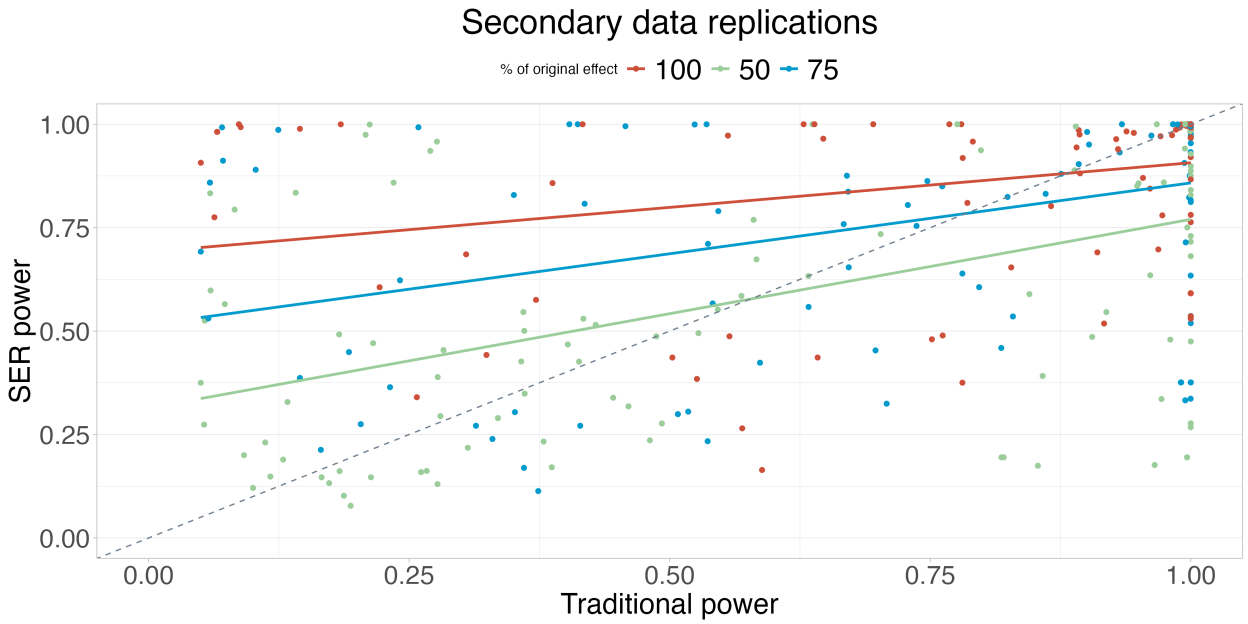


Figure S4. New data replication power estimates for SER versus traditional approach.

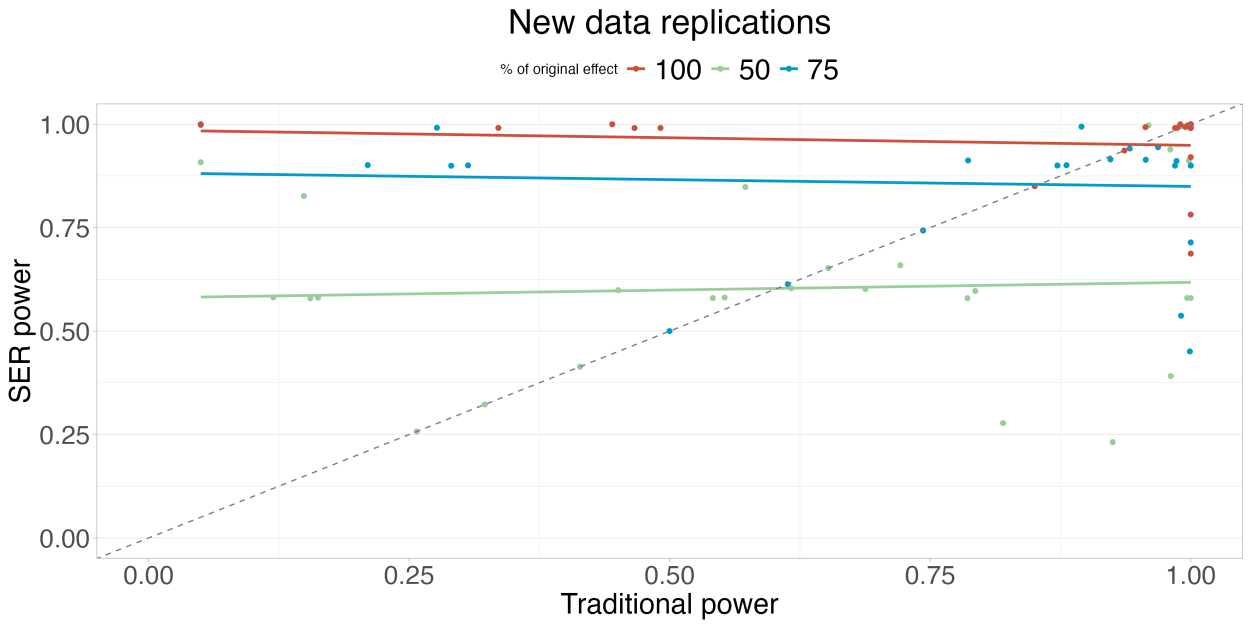


Figure S5. Distribution of power estimates across traditional, SER, and combined approaches.

Onboarding Replication Teams

Once a team was matched, project coordinators sent an onboarding email that also included a presentation with all pertinent information provided. The team was invited to an onboarding call to relay key information and answer questions. A project coordinator would reach out to ensure the researchers' institution had an active Federalwide Assurance (FWA), a requirement to participate in SCORE, and that the team could complete the replication with available resources and time. Primary feasibility criteria was average length of IRB review at the team's institution, availability of appropriate expertise and instrumentation, availability of the required sample, and appropriateness of the proposed budget. If the team could complete the study within the project timeline, coordinators created a unique ID number for the replication, then created and shared an OSF project, a preregistration template (as a Google document), and IRB/HRPO instructions with the researchers.

The coordination team provided training to replication teams for using OSF and adhering to preregistration and documentation standards for the project. Statistical consultants provided just-in-time support for statistical and methodological issues during the preparation of preregistrations to eliminate any time bottlenecks.

Once a lab showed interest in performing a project on a paper, they completed a budget proposal. Human subjects data collection projects were capped at a maximum of US$10,000 per project, and collaborating labs were asked to specify how much money they were requesting to complete their project and how that money would be allocated between personnel costs, participant recruitment, and any materials needed. These budgets were reviewed and approved by the coordinating team in accordance with guidelines set by the project funding agency.

Recruited replication teams, and also reviewers and editors, were given the following instruction for helping them to think about how to design, evaluate, and conduct their replications: "*Remember that your goal in replication is to achieve a design that is a good faith test of the original claim [linked to a preprint of Nosek & Errington, 2020]. Sometimes that is a*

straightforward repeat of the original procedure in your new sample. Sometimes that means adapting the methodology for the new context. Changing the methodology is not bad if it is done so in the service of improving the quality of the replication for testing the original claim. Also, so that we do not introduce new challenges or delays, keep your goal very focused on testing the original claim and not introducing new design elements that could create complications in IRB or preregistration review."

This instruction embraced the perspective that deciding something is a replication is a theoretical commitment (Nosek & Errington, 2020), and the researchers, reviewers, and editors task was to assess how best to conduct the replication to be a good-faith attempt to replicate the original claim. This can mean using the procedures exactly as was done in the original study, or it can mean adapting the procedures given the changes to the context of the new study given the theoretical understanding of the original claim.

This differs from conceptions of replication that are made purely in procedural terms -- assessing how much the replication procedures differ from the original procedures. When defined in procedural terms exclusively, the concept of "how much" is not meaningful from the perspective adopted for this investigation. Take the example from the main text:

Conceptually, it can be challenging to attempt a replication of a prior finding. There is no such thing as exact replication. Replications inevitably differ in many ways including the units, treatments, observations, and settings from the original research. Researchers must make decisions about how to conduct a good faith replication of an original claim. For example, should a present day replication of a 2009 U.S. study of political behavior that used President Obama as a stimulus use Obama again, use the current U.S. president, or use the leader of the participants' nation? The answer depends on what features of that stimulus are essential for testing the original claim. The decision that a new study is a replication of a prior study is a theoretical commitment that they are testing the same claim.

Imagine a study that kept everything else the same and only changed the stimulus (Obama) to be the current president (Trump). This is a small procedural change, but it is not at all clear if it is a small or large change to test the original claim. For example, if the research question were about how people respond to biracial leaders, then this small procedural change is a huge change to the design, given the research question. Trump is not biracial, this small procedural change would disqualify the study as a good-faith replication.

To offer another example. Imagine an original procedure that involved administering a survey was repeated exactly in a new sample, with no changes. Is that a close replication? It could be. But, what if the new sample were 4 year-olds who can't read, or English-only readers when the original survey was in German? For any sensible research question, administering the identical procedures to those samples would be considered a huge change--not a close replication-- despite the fact that nothing was changed in the procedures themselves. The fact that everyone easily recognizes that the procedures need to be adapted in this example is because the auxiliary theoretical assumptions are so embedded as to be barely noticeable -- the respondents need to be able to understand the survey questions to be able to respond to them. This auxiliary assumption is not part of the procedures, it is part of the conceptual understanding of the research aims. In sum, the notion of "how much" is a theoretical assessment based on the relationship between the research question and the study design, not a feature of the study procedures on their own.

So, how did we manage this? In this investigation, all studies need to meet the criterion of "good faith replication" through the internal peer review process. Replicators, reviewers, and editors

were assessing whether the design -- with whatever changed or did not change -- met the theoretical standard of being a good faith replication. That is, all replications should be “close” replications appropriate for testing the same question -- close being a theoretical assessment, not a procedural one. The concept of conducting a good faith replication and its relation to procedural characteristics of studies was developed in depth by Nosek and Errington (2020) to which we referred replication teams, reviewers, and editors. To highlight one relevant long quote from that article (p. 2):

“*The repetition of the study procedures is an appealing definition of replication because it often corresponds to what researchers do when conducting a replication—i.e., faithfully follow the original methods and procedures as closely as possible. But the reason for doing so is not because repeating procedures defines replication. Replications often repeat procedures because theories are too vague and methods too poorly understood to productively conduct replications and advance theoretical understanding otherwise.*

Prior commentators have drawn distinctions between types of replication such as “direct” versus “conceptual” replication and argue in favor of valuing one over the other. By contrast, we argue that distinctions between “direct” [or “close”] and “conceptual” are at least irrelevant and possibly counterproductive for understanding replication and its role in advancing knowledge. Procedural definitions of replication are masks for underdeveloped theoretical expectations, and “conceptual replications” as they are identified in practice often fail to meet the criteria we develop here and deem essential for a test to qualify as a replication.

We propose an alternative definition for replication that is more inclusive of all research and more relevant for the role of replication in advancing knowledge. Replication is a study for which any outcome would be considered diagnostic evidence about a claim from prior research. This definition reduces emphasis on operational characteristics of the study and increases emphasis on the interpretation of possible outcomes.

To be a replication, 2 things must be true: outcomes consistent with a prior claim would increase confidence in the claim, and outcomes inconsistent with a prior claim would decrease confidence in the claim. The symmetry promotes replication as a mechanism for confronting prior claims with new evidence. Therefore, declaring that a study is a replication is a theoretical commitment. Replication provides the opportunity to test whether existing theories, hypotheses, or models are able to predict outcomes that have not yet been observed. Successful replications increase confidence in those models; unsuccessful replications decrease confidence and spur theoretical innovation to improve or discard the model. This does not imply that the magnitude of belief change is symmetrical for “successes” and “failures.” Prior and existing evidence inform the extent to which replication outcomes alter beliefs. However, as a theoretical commitment, replication does imply precommitment to taking all outcomes seriously.

Preregistration and Peer Review

The preregistration review process was designed with the goals of promoting integrity, transparency, and rigor for the replication studies. We subjected each replication study's design and analysis plan to peer review before conducting the research. The approach mandated replication teams to engage in open science practices throughout the project. Ideally, these practices would promote accountability for the research team, reproducibility of the outcomes, and accessibility of the research process and outputs to enable examination by interested readers.

For each replication study, teams were required to articulate the claim to be evaluated, outline their study design and sampling plan, specify variables of interest, and describe their analysis plan before conducting any analyses. Teams were provided standardized formats for documentation and protocol planning. These protocols were scrutinized by peer reviewers, who evaluated the strength of the proposed methods and appropriateness of the design for testing the same question as the original study. The peer review process was managed by editors who approved when the replication study was ready to be conducted. Functionally, the process mimicked the Stage 1 peer review process for Registered Reports (Chambers, 2019; Chambers & Tzavella, 2022), but was conducted in a peer review system set up and customized for the purposes of the program rather than in association with a journal.

Recruiting Reviewers and Editors

During recruiting for collaborators, researchers could indicate interest in conducting replication or reproduction studies, and also potential interest in serving in editorial or reviewer roles. Program leaders conducted personal outreach to researchers with substantial experience in editorial roles at disciplinary journals across the social-behavioral sciences to participate as editors for SCORE. Table S2 identifies the Editors that managed the peer review process for one or more replication studies. Replication projects for which they served as editor are identified by their OSF ID which can be found by replacing “abcde” with the five character ID in the following link: https://osf.io/abcde.

Table S2. Editors for SCORE replication studies peer review process.

Name	Institution	Title	OSF Links
Amélie Gourdon-Kanhukamwe	Kingston University	Lecturer	vcfsg, apv5t, 5du7f, jxd7b, rgtnm, f5b6c, yq3jb, hz97d, rcv49, 9n8uh, nb3z9, u52y3, 834de
Annette Brown	FHI 360 (Strategy and Innovation)	Head of Strategy and Principal Economist	na3mh, 3tczs, t9h6p, fhyzs, z4nyv, 8vn2z, agv64, sv4e8, 5tk2e, ta5vh, kgqu5, m4yse, zt5y2
Bert Bakker	University of Amsterdam (Amsterdam School of Communication Research)	Associate Professor	dbw54
Bobbie Spellman	University of Virginia (School of Law)	Professor	md84s, rqwju, ba26p
Cristina Zogmaister	Università di Milano-Bicocca (Dipartimento di Psicologia)	Associate Professor	cs8y2
M. Donabel Yap	Freelance Consultant	Consultant	68x45, vhu2x, 9vxwz, k6xm9, h7av9, bmzuf
EJ Wagenmakers	University of Amsterdam (Psychology)	Professor	wtm3e, wdgns, tdr6s, w8xun
Eric Sevigny	Georgia State University (Criminal Justice & Criminology)	Professor	8ndcx
Ernest O'Boyle	Indiana University (Management and Entrepreneurship)	Professor of Management	dk3xz, jsy47, hq4cr, q4fkd, s5v28, eumv2, 6vhnz, xtmy9, n2kma, xzufr, r924g, ty9du, kvsa6, 2dx54, afznv, jb9rq, 4v7ax, devn8, 9u2n4, 4h6v2
Gustav Nilsonne	Karolinska Institutet (Department of Clinical Neuroscience)	Associate Professor	m36y2, 9qjhf, 62xh4, kr7wf
Heather Kappes	The London School of Economics and Political Science (Management)	Associate Professor	qtv6j, 8qkrx, e9y6b, e52qx, vngz9, 2jn6d, 8eaqp
Jan Roer	Witten/Herdecke University (Department of Psychology and Psychotherapy)	Professor	23ghe, wjh3c, 5k6jg, ujg2c, 3g74a, ebaf2, skj4c, 7vty6, 25mvt, h7wx8, 2be7n, utjhd, 7f42k, q7hpx, yg7d2, u5r47, 7k8sy, q7jk4, z9emg, vg6zy, 8qujb, h37zx
John Lloyd	University of Virginia (Education and Human Development)	Professor Emeritus	2cjuw, 9r2jw, e7tw2, a4swd, mu4rs, bj9ea, 87qs6, 7qnrs, qczeu, bmwsv, en9md, kr7wf
Kai Jonas	Maastricht University (Work and Social Psychology)	Professor	p7mex, cwmb5, n35c2, ty5be, exrua, s2x3c, tuy3f, vt6ga, tmn3q, bx4hy, p2st5, mztkr, c9n5x, vcjf6, uwnya, e48gd, tz9ef, 697dv, em34f
Kevin Esterling	University of California Riverside (School of Public Policy)	Professor	8uc5z, rvs4q, 2ug8x, rpumq, rs7e9, wy3tj, 8pxty, xtq68, 4w5g2, nse8t, 8pxty
Kim Peters	University of Exeter (Management)	Professor	ftbmk, 953fa, fncqp, g9v87, y47xc, w56ve, jdb7q, j8en2, 925nz
Kimberly Quinn	DePaul University (Department of Psychology)	Professor	56jqu, h24z8, 8feku
Miguel Fonseca	University of Exeter (Business School)	Associate Professor	h2cq7, qrjsu, p9k76, fgx67, q4fpr, 82g9j
Nathaniel Porter	Virginia Tech (University Libraries)	Assistant Professor	zfxk9, 6aynt, 4rjbf, u8x9s, efuhg, qru6t, er54t, usv5k, 9mx5w, rywgh, c6pkb, guen5, w2xrp, shvdk, qgm2c, n7m39, qxk38, bcqem, gk92w, psmq5, k4suw, df36m, jegqs
Nick Fox	Center for Open Science	Research Scientist	qy62r
Nick Weller	University of California, Riverside (Political Science)	Associate Professor	5fjd4, 7tc8b

Richard Lucas	Michigan State University (Psychology)	Professor	etfqz, er4gy, npju2, aykfp, a8eg9, uyt9p, mp3ak, cr6zm, hd3vf, huavw, snx3t, tacvb, e6mfw, 9nt6s, su48v, a6ksz, n9vkh, q8bdv, xnh2r
Rolf Zwaan	Erasmus University Rotterdam (Department of Psychology, Education, and Child Studies)	Professor	78fvn, 5qwdh
Simine Vazire	University of Melbourne (Melbourne School of Psychological Sciences)	Professor	7gcjf, dy3wx, sw67c, 8hp5j, tvhma, qe8sp, dhg3q, jhaur, rjhwt
Tabare Capitan	Swedish University of Agricultural Sciences (Department of Economics)	Postdoctoral Researcher	b48am, 6aynt, nd87t, bawh3, t95k2, 2dzn7, yv2am, ub4c6, h8ecm

We invited researchers to serve as independent reviewers of others' projects. To express interest, researchers filled out a short form, identified the journals that published research in their expertise area, and provided their CV. The process of selecting reviewers started as a centralized process with coordinators matching reviewers and editors to protocols. Later in the process, we shifted to sharing protocols that were available for review by email to the reviewer pool and asking reviewers to self-identify interest, expertise, and availability. Reviewers and editors were paid per review unless they declined payment.

Drafting preregistrations with the design and analysis plan

Replication teams described their research plan using a New Data Preregistration Form or a Secondary Data Preregistration Form. The specific claim and sample size were provided by the coordinating team to the replication team. The preregistration forms were based on the standard OSF preregistration template. They included SCORE-specific instructions to guide a researcher through each step. Also, the coordination team provided guidance and answered questions as needed.

For secondary data replications, teams were split into two roles during Phase 1: *data finder* and *data analyst*. A data finder identified available data sources that could be used to conduct high-powered and good-faith replications of the selected claims and prepared the dataset for analysis. The coordinating team evaluated prepared datasets to ensure that they would allow for a high quality test of the original claim. Upon approval from the coordinating team, the *data finder* completed sections of the preregistration that offered procedural information about the data collection process, including the steps required to access the data, an outline of the variables necessary, a detailed narrative describing how the various datasets were cleaned and merged into a final replication dataset along with a commented analysis script (if available), and a data dictionary. The preregistration document was then shared with the *data analyst* along with the paper, the claim to be replicated, all of the materials developed by the *data finder*, any available code provided by the original author, and a portion of the replication dataset. The *data analyst* completed the remaining sections of the preregistration, including description of the analysis plan, and provided final replication outcomes. For the secondary replications conducted during Phase 2, the data finder and data analyst roles were combined into a single person or team.

Peer review process

Once a draft of a preregistration was complete, it was reviewed by between 1 and 3 reviewers. An Editor oversaw the review of each paper, and resolved any conflicts regarding design that were not resolved during the review process. To facilitate easy and asynchronous collaboration, each review was conducted in Google Docs. The reviewers' questions and suggestions were posed using the comment feature. Reviewers reviewed the research designs for clarity,

completeness, and quality. Reviewers did not evaluate the quality of the original research, plausibility of the findings, or advisability of the design decisions that are appropriately faithful to the original methodology and findings because the purpose was to conduct a good-faith replication of the original research, regardless of its merits. The review period was set to five days, after which access to the document was closed to external participants and the replication team addressed all outstanding reviewer and editor comments. Often, issues raised by reviewers took additional time to address beyond the five-day window. On average, reviews lasted 7.5 days for new data replications and 19.5 days for secondary data replications.

The Editor assessed whether the replication team adequately addressed the comments, and either asked for additional revisions, approved, or rejected the replication. Once approved by the Editor, the final draft of the preregistration was added as a word document or PDF to the OSF component designated for this replication study and formally registered using the OSF Preregistration template. The coordinating team documented the key variables expected at the conclusion of the replication study, and the replication and coordinating teams aligned their knowledge of what materials were to be delivered at the conclusion of the study. After this, replication teams formally registered their preregistration and any associated materials on the OSF project component they would use for the management of project-relevant files, data, and outcome reporting.

Upon acceptance of the preregistration, teams received a $500 award for successfully completing the preregistration, and an additional $100 if they had submitted the preregistration for review within a week of being matched with a project to incentivize prioritizing the replication.

Engaging original authors

For new data replications, original authors were invited to engage during the peer review process. Just like other reviewers, they could make comments and suggestions directly in the preregistration Google doc. After a preregistration was approved by the Editor, original authors were invited to submit additional feedback as commentary. To ensure transparency, commentaries were uploaded to the replication OSF project alongside the final preregistration.

For secondary data replications, we decided not to include original authors during the peer review process because the data was often publicly accessible and may have introduced unwelcome temptation for original authors to test their original claims during review. Instead, the original authors were provided with a link to the final preregistration after the review period and invited to submit a commentary following the same approach with new data replications.

Acceptance rate

During Phase 1, a total of 125 new data replication plans and 73 secondary data replication plans were peer reviewed. Because the goal was to conduct good-faith replications of as many claims as possible, we tried to support revisions and improvements to make all replications acceptable. However, 4 new data replication plans and 7 secondary data replication plans were ultimately rejected by an Editor, or a 97% and 90% acceptance rate respectively. In these cases, Editors cited unresolvable issues with the study designs or data sources that would have prevented these studies from being high-quality, good faith verifications of the original claim.

During Phase 2, a total of 35 new data replication plans and 50 secondary data replication plans were peer reviewed. Of those, only one secondary data replication plan was ultimately rejected by an Editor (100% acceptance rate for new data replications; 98% acceptance rate for secondary data replications).

Replication inferential criteria

Criterion for a successful replication attempt within the SCORE program was achieving a statistically significant effect (alpha = .05, two tailed) in the same pattern as the original study on the focal statistical evidence. This criterion was used to judge whether humans and machines could predict replication outcomes. We also estimated replication success with other binary assessments methods and effect sizes. To the extent possible, we transformed effect size estimates to a standard metric for reporting in the aggregate across replications. A section below provides additional details about effect size conversions.

For the 200 “bushel” papers in which multiple claims were extracted from a single paper, if there was more than one significant test result for a claim, the test judged to be the most central one was used for the purposes of inferential evidence. If there were multiple test results in the original study, but only one was a significant inferential test, then that was chosen for the inference criteria. And, if there were no significant inferences in the original claim, then the replication outcome was not included in the dataset per the inclusion requirement for a significant result.

The main text provides an overview of 13 binary assessments of replication success. Below are additional details for a few of those assessments.

Subjective Interpretation

The subjective assessment of replication success was made by replication teams or by project coordinators reviewing the submitted outcomes for the replications. There was no explicit guidance for how to make the subjective assessment other than to evaluate whether the claim had replicated successfully. Raters could indicate: Yes, No, or Complicated. In Phase 2, an option of “Not Sure” was added, though none of the replication outcomes reported here were rated as such. For binary assessment, only “Yes” was considered replication success and cases of “Complicated” were excluded.

Meta-analysis

We synthesized the original and replication studies by means of the fixed-effect meta-analysis model (Hedges & Vevea, 1998; also known as the common-effect or equal-effect model in other implementations). This model does not assume that the true effect size underlying both studies is the same, nor does it allow generalizing the results to a population of studies. Instead, it is concerned with reliably combining evidence from two independent studies. While equal-effect and fixed-effect models are conceptually distinct, in this implementation the outputs of the models are identical (Laird & Mosteller, 1990; see also https://wviechtb.github.io/metafor/reference/misc-models.html).

Bayesian meta-analysis

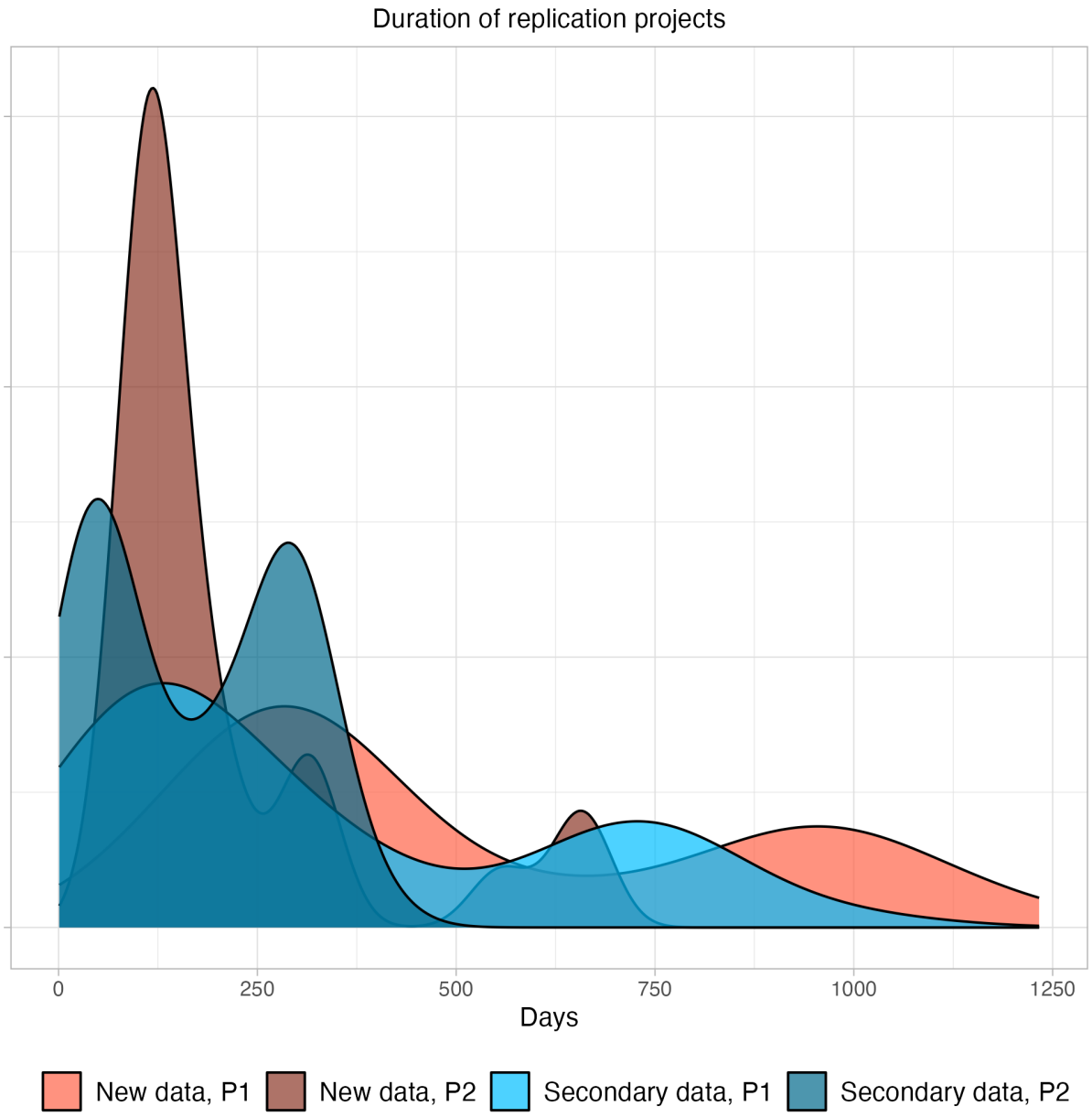
We used a fixed-effect model for this analysis to match the approach used immediately above. In our implementation, (https://danheck.github.io/metaBMA/articles/metaBMA.html), we estimated the model with a prior on the effect size centered at 0 with a standard deviation of 0.25.

Data collection & analysis

Replication teams completed the protocols with active monitoring and assessment by the Project coordinators throughout the process to ensure adherence to quality control expectations

and timeline. Figure S6 shows the distribution of time required to complete replication studies separated by phase and whether the replication involved new data collection or used secondary data. Means and standard deviations of project duration are as follows: Phase 1 new data = 517 days (332 days), Phase 1 secondary data = 339 (284), Phase 2 new data = 247 (171), Phase 2 secondary data = 168 (122).

Figure S6. Number of days to complete replication projects from the date the first researcher was added to the study's OSF project to the date the final report was last modified.



Caption: P1 = Phase 1; P2 = Phase 2. The completion date for replication studies was hardcapped at March 31, 2023 to reflect the late date of SCORE data collection.

Most SCORE replications were conducted after the beginning of the COVID-19 pandemic in Spring 2020. Because of the pandemic seven of the preregistered replication attempts were not

completed, either because data collection was not possible or because pandemic-related delays made the replications infeasible with the SCORE timeline. Additionally, in 13 completed studies data collection was shifted online (either entirely or in part); in four studies data collection was cut short or the pandemic was thought to otherwise depress the sample size; in one study the in-person data collection procedure was altered; and in two studies the replication researchers explicitly flagged that the pandemic could have substantively affected the replication attempt.

Outcome reporting

Replication teams authored reports of their observed results and a comparison with the original study. A reporting manual and template guided replication teams through the outcome reporting process. These included instructions for uploading relevant files to the respective OSF project. Replication teams also filled out an outcome variable form to report key information from the study in a structured format. Once the replication variables were returned, a coordinating team member and statistical consultant would assess the reporting and calculate missing variables from original studies, replication studies, and those used to evaluate whether replications were successful. Those results were then reported in a standardized format for extraction to the database and for referencing to the study code and data. A coordinating team member would then evaluate the written report for completeness and accuracy before accepting them and denoting the project as complete. They would also review that private data was removed from any components that would be made public. A complete project included project content (IRBs, materials, data, scripts, etc.) being posted to the project on OSF in a standardized format.

Table S3 provides links to all OSF projects for replication attempts that were started. Paper ID and Project ID columns provide the project-specific identifiers used for tracking and project management. For any given project, replace “abcde” in the link https://osf.io/abcde with the five characters in the OSF column to find the plans, materials, data, and reporting on OSF. The “data collected” column is marked yes if at least some data was collected for that replication attempt. The “completed and reported” column is marked yes if the project met inclusion criteria and outcomes are reported in this paper.

Table S3. Identifiers and links to OSF projects for all replication attempts started

Paper ID	Project ID	OSF	Data Collected	Completed and Reported
0PZl	41y2	5chvj	No	No
0qar	20g	w8xun	Yes	Yes
0wWk	28yg6	efuhg	Yes	Yes
1574	2w9k2	mztkr	Yes	Yes
1574	4100	k9geu	No	No
1VeW	g941	3tczs	Yes	Yes
1X9W	32mk	9qjhf	Yes	Yes
25Yg	97g	2qyfr	No	No
2GKO	4142	nse8t	Yes	Yes
2Ypg	28gg6	dq4k9	No	No
2lb5	y496	q4fpr	Yes	Yes
347d	m89	56jqu	Yes	Yes
3aPw	05g8	qczeu	Yes	Yes
3aPw	2y4gm	k4suw	Yes	No
4q0L	3z5z	m4yse	Yes	Yes
521q	286	aykfp	Yes	Yes
598w	79k5	kwsdr	No	No
59bm	2196	xd9c8	Yes	Yes
59bm	m4k7	953fa	Yes	Yes
59bm	z96	y47xc	Yes	Yes
5Awm	2y2g	k6xm9	Yes	Yes
5JE	z189	agv64	Yes	Yes
5Kgq	k6y7	62xh4	Yes	Yes
5KrD	935y	sv4e8	Yes	Yes
5Q82	6zy82	guen5	Yes	Yes
5XEE	21z56	n35c2	Yes	Yes
5Xaw	69y36	7vty6	Yes	Yes
5Xeq	8	wu5pv	No	No
7RR2	86m96	q7hpx	Yes	Yes
7RR2	0y08	evrxs	No	No
7WjP	m7y9	d8we7	No	No
7WjP	7976	md84s	Yes	Yes
7X54	93k7	7qnrs	Yes	Yes
7ybJ	z2416	w2ahm	No	No
88xa	1554	afznv	Yes	Yes
8R9d	9kgg	mp3ak	Yes	Yes
8qER	56z6	6ycdg	No	No
8w97	gz2m	sw67c	Yes	Yes
8wZ0	276	tdr6s	Yes	Yes
8wZ0	2y486	vt6ga	Yes	Yes
9DZl	548	9s4tw	No	No
9J1	g9mm	rqwju	Yes	Yes
9OK1	658g7	h8ecm	Yes	Yes
9R9X	2kkg2	tuy3f	Yes	Yes
9Yqy	99yg	xh4eq	No	No
9ey	z161	ftbmk	Yes	Yes
9wkl	94ky	6aynt	Yes	Yes
9wya	9k2y	4w5g2	Yes	Yes
AOQj	21487	697dv	Yes	Yes

AYQG	6m33m	n7m39	Yes	Yes
AYQG	6g7k	cfra4	No	No
AgO1	895g	5tk2e	Yes	Yes
AgO1	23m12	z9emg	Yes	Yes
AqDO	2wyw2	w2xrp	Yes	Yes
AqDO	7945	tacvb	Yes	Yes
AvOr	24716	bx4hy	Yes	Yes
AvOr	92g	jxd7b	Yes	Yes
AvWY	8g91	uyt9p	Yes	Yes
BKxK	4zz0	nd87t	Yes	Yes
BaDx	0m7	4unjb	No	No
BIRQ	546	kr7wf	Yes	Yes
Blxd	6z3o2	em34f	Yes	Yes
Blxd	3053	mu4rs	Yes	Yes
Br0x	658m2	tmn3q	Yes	Yes
Br0x	7g66	n2kma	Yes	Yes
Br0x	2kgg2	25mvt	Yes	Yes
BrGp	247z3	h37zx	Yes	Yes
D2LY	65o96	qgm2c	Yes	Yes
DEqr	61k8	8uc5z	Yes	Yes
E4Am	m5y9	n9vkh	Yes	Yes
E5qr	65gm6	shvdk	Yes	Yes
EAa	yz66	yhtz4	No	No
EJpm	288ok	gk92w	Yes	Yes
EJpm	yk20	7tc8b	Yes	Yes
EJpm	5738	7zn4v	No	No
EKBZ	618k	e9y6b	Yes	Yes
EQxa	6738	wdgns	Yes	Yes
EQxa	6m396	7f42k	Yes	Yes
EZ3x	8z2g	a4swd	Yes	Yes
Eb2N	23yw2	rywgh	Yes	Yes
Ej3y	65km6	skj4c	Yes	Yes
G1Lr	gz9m	mjfwp	Yes	Yes
G1Lr	887	jsy47	Yes	Yes
G1Lr	387	6vhnz	Yes	Yes
G1Lr	369z	eumv2	Yes	Yes
G1Lr	999g	r924g	Yes	Yes
GNjz	y60	m36y2	Yes	Yes
GXEW	67g8	834de	Yes	Yes
J40k	67z8	4h6v2	Yes	Yes
J4W9	k144	qrjsu	Yes	Yes
J4W9	y2312	yg7d2	Yes	Yes
J7Z2	317z	x9f3m	No	No
J7ek	89z7	yq3jb	Yes	Yes
J7ek	288g2	exrua	Yes	Yes
JRpA	2g7ky	bcqem	Yes	Yes
JRpA	mk67	xtq68	Yes	Yes
JWzJ	6547	u8x9s	Yes	Yes
K4ZD	y071	2jn6d	Yes	Yes
Kj9d	16yz	5du7f	Yes	Yes

Kybl	23w8w	bmzuf	Yes	Yes
LbEB	6zzo6	ty5be	Yes	Yes
LbEB	gz7m	ewujm	No	No
LmBx	y2gz6	p2st5	Yes	Yes
LyLd	944y	jb9rq	Yes	Yes
LyLd	10g2	9u2n4	Yes	Yes
N8pB	2ggz2	ujg2c	Yes	Yes
Nj8V	746	2dx54	Yes	Yes
Njqj	9k67	96vn7	No	No
NrrW	24812	9vxwz	Yes	Yes
OKRy	281g6	5k6jg	Yes	Yes
OYX0	658y7	yv2am	Yes	Yes
OYX0	8zz7	bj9ea	Yes	Yes
OeGv	24m16	as7te	Yes	Yes
OeGv	y050	8pxty	Yes	Yes
Ovkm	zz26	f5b6c	Yes	Yes
Ow0	46z8	q8bdv	Yes	Yes
P1rY	99kg	ba26p	Yes	Yes
P8Ab	2wmk2	6q87t	No	No
P9Vr	24z12	tj9hy	No	No
PNPz	5zg9	er4gy	Yes	Yes
Pb9K	7g95	rvs4q	Yes	Yes
PlDa	786	huavw	Yes	Yes
Q1dl	40z0	ndbrp	Yes	Yes
Q1dl	om7	hq4cr	Yes	Yes
Q1dl	1642	xzufr	Yes	Yes
Q1dl	g6936	3b49u	No	No
QYNq	kk47	5fjd4	Yes	Yes
Qk5P	3zz	7rtpw	No	No
QllV	249w6	bawh3	Yes	Yes
R8RN	1yz2	kvsa6	Yes	Yes
R9dv	42m8	z4nyv	Yes	Yes
RJb	6m4k	vngz9	Yes	Yes
RYKv	2k5g2	df36m	Yes	Yes
RYKv	mkz7	ta5vh	Yes	Yes
RZdL	6o946	c9n5x	Yes	Yes
RrOb	19gz	devn8	Yes	Yes
Rvb	2y582	ebaf2	Yes	Yes
VB9K	2g7z2	vg6zy	Yes	Yes
VGAe	4z32	dk3xz	Yes	Yes
VOwm	ykk0	82g9j	Yes	Yes
Vj0p	2g79y	tz9ef	Yes	Yes
Vj0p	yzm6	e6mfw	Yes	Yes
VjYA	y01	kdr3f	No	No
VjjX	99m7	9r2jw	Yes	Yes
Vpgm	m7m3	hz97d	Yes	Yes
VvZX	4182	p9k76	Yes	Yes
VvlD	2616	8vn2z	Yes	Yes
WLkV	67516	q7jk4	Yes	Yes
WLpV	1964	t9h6p	Yes	Yes

WaYe	556	fgx67	Yes	Yes
Wre	4130	h24z8	Yes	Yes
Y8Yx	21wg2	usv5k	Yes	Yes
YWep	75g6	u52y3	Yes	Yes
YmQR	69412	c6pkb	Yes	Yes
YpZZ	4930	g6fkp	Yes	Yes
YpZZ	3z7z	rjhwt	Yes	Yes
YpZZ	15yz	qe8sp	Yes	Yes
YpZZ	8y1	jhaur	Yes	Yes
YpZZ	4zy8	7gcjf	Yes	Yes
a8jQ	6mz92	3g74a	Yes	Yes
aYyR	23w12	7k8sy	Yes	Yes
aag9	6z1o2	wjh3c	Yes	Yes
amYY	6mz8	h2cq7	Yes	Yes
bLe8	329k	t95k2	Yes	Yes
bY2A	6zoo2	vhu2x	Yes	Yes
d24p	67z12	cwmb5	Yes	Yes
d5v3	6m492	vcjf6	Yes	Yes
dLKV	11z	9mb25	No	No
dqKX	7965	e7tw2	Yes	Yes
dxQp	2w5k2	qru6t	Yes	Yes
e227	24yw6	9mx5w	Yes	Yes
e3G7	6o442	23ghe	Yes	Yes
eOQm	yk91	2cjuw	Yes	Yes
eRbK	468	qf4c2	No	No
eg1q	6g67	8hp5j	Yes	Yes
eg3p	yyy1	e52qx	Yes	Yes
egX9	4z88	fncqp	Yes	Yes
exBp	2gy82	jdz79	No	No
exd7	318z	qtv6j	Yes	Yes
gRWz	m5k3	tvhma	Yes	Yes
gbAY	g7g1	8eaqp	Yes	Yes
gbAY	165m6	8qujb	Yes	Yes
gbg4	8z7g	npju2	Yes	Yes
gbl9	393	vcfsg	Yes	Yes
j2yd	5796	su48v	Yes	Yes
j2yd	m5m7	sxqt8	No	No
j2yd	1994	sxca4	No	No
jDWN	0y38	g9v87	Yes	Yes
jLr	4z02	dy3wx	Yes	Yes
jLr	m707	dhg3q	Yes	Yes
kXp8	kzyz	925nz	Yes	Yes
lJ0w	931g	ty9du	Yes	Yes
ld8V	y50	2afjs	No	No
lkBL	618	cr6zm	Yes	Yes
mBL1	8m17	8qkrx	Yes	Yes
mJDj	m7g7	snx3t	Yes	Yes
mJDj	g79m	9nt6s	Yes	Yes
mJDj	y7g1	u5d2y	Yes	Yes
mrZ	579	na3mh	Yes	Yes

mxyQ	g77m	2ug8x	Yes	Yes
pN7E	69812	f49uq	Yes	Yes
plLK	5g9	fhyzs	Yes	Yes
plLK	6ow46	h7wx8	Yes	Yes
ppz8	4968	4cgdw	No	No
pw3m	2kwg2	68x45	Yes	Yes
q8xv	mkk9	jdb7q	Yes	Yes
q8xv	23g12	ub4c6	Yes	Yes
qPxQ	69136	2be7n	Yes	Yes
qQ9Z	y71	vdezq	No	No
qXX2	gz51	zt5y2	Yes	Yes
qYr7	675m9	qxk38	Yes	Yes
qYr7	5z36	wy3tj	Yes	Yes
qgWj	2yg	kgqu5	Yes	Yes
qgWj	23wwg	xwthk	Yes	Yes
rjb	9kzy	rgtnm	Yes	Yes
rjb	24316	utjhd	Yes	Yes
rpq	308	hd3vf	Yes	Yes
rvyb	356	7tr8a	No	No
vGqL	67y16	er54t	Yes	Yes
vkYO	109z	8weqr	No	No
w5dv	657	q4fkd	Yes	Yes
wRvv	65wm6	e48gd	Yes	Yes
wRvv	1634	rpumq	Yes	Yes
x0pA	8z81	en9md	Yes	Yes
x2XO	1y22	bn6tj	No	No
x3KP	k5z	a8eg9	Yes	Yes
xGGO	y11	w56ve	Yes	Yes
xYbO	yk16	8feku	Yes	Yes
xvrb	z591	xnh2r	Yes	Yes
y3R4	2wgk2	p7mex	Yes	Yes
yAPR	1762	j8en2	Yes	Yes
yDyG	z106	xtmy9	Yes	Yes
yJwG	9ky	wtm3e	Yes	Yes
yQeR	38	4v7ax	Yes	Yes
yypJ	6g38	apv5t	Yes	Yes
z0v1	yz21	57f2j	No	No
z4dO	yzz0	rs7e9	Yes	Yes
zK2	m5g9	etfqz	Yes	Yes
zN22	m63	5qwdh	Yes	Yes
zNEm	1y02	78fvn	Yes	Yes
zb3Y	2y182	h7av9	Yes	Yes
zb3Y	41k2	bmwsv	Yes	Yes
zekm	234w6	yfmg8	Yes	Yes
zekm	k17	s5v28	Yes	Yes
zlBL	651m6	uwnya	Yes	Yes
zlBL	z516	rcv49	Yes	Yes
zlm2	9977	qy62r	Yes	Yes
zlm2	128g6	s2x3c	Yes	Yes
zlw	wz9	jcs54	No	No

zmYY	g5m	a6ksz	Yes	Yes
zmYY	6zz3k	8ndcx	Yes	Yes
zqwm	z4z9	87qs6	Yes	Yes

A key part of the replication design process was peer review in advance to maximize the chances that designs were good-faith attempts to replicate an original finding. We sought to document decisions about research designs transparently so that differences between replication designs and original studies could be made visible to support future hypothesizing about why similar or different results were observed. Moreover, we sought to document differences between the preregistered replication designs and what ultimately occurred when carrying out the research as sometimes changes are made because of unavoidable circumstances (e.g., COVID pandemic) or because problems with the original design are identified and corrected during execution. Replication teams preregistered their study designs and provided final reports that included sections to document deviations. Table S4 (deviations from original study) and Table S5 (deviations from preregistration) extract the deviations reported by replication teams from their reports. We first used an LLM to prepopulate the tables with extracted deviation information from each report. We then did a manual check to identify errant inclusions and exclusions. Projects with no relevant text identified are excluded from the Tables. In both tables, the reports can be accessed by replacing “ABCDE” in https://osf.io/ABCDE with the five characters in the OSF ID column of each table. Those reports provide fuller context of the replication designs to assess the rationale and potential consequences of the reported deviations, as well as content that was not identified in the automated text extraction or the manual review used to construct these summary tables.

Table S4. Deviations in replication designs compared with original studies reported by replication teams

Project ID	OSF ID	Reported deviations from original study
99kg	tb2up	Deviations from the original study: Pickett and Baker originally used Survey Monkey's "online Audience panel" to recruit participants. Participants were Americans aged 18 or older (self-report). They found that "non- Whites and persons without a college degree are underrepresented, which is normal in Internet surveys." Their sample was 78% White and averaged 3.83 out of 5 on education (between some college and college degree). Pickett and Baker did not provide the frequencies of each educational category. The current study was 63% White and averaged 3.42 out of 5 for education with 3.7% less than high school, 21.4% high school degree, 24.9% some college, 29% college degree, and 21.1% with a graduate degree.
y60	n8tds	Deviations from the Original Study Participants were recruited from Coventry University rather than the University of Stirling, and they were offered research credits where no reward was previously offered. It is not expected that this would impact the effect of interest as they likely represent a similar UK student population, but was adopted to maximise the likelihood of meeting the data collection targets. Different images were used as stimuli due to the lack of access to original materials (see Stimuli Development document on the OSF for noting of deviations in stimuli development). The result is a collection of stimuli which are slightly more obviously manipulated than the original however nevertheless replicate the original manipulation process and represent relatively realistic images for participants to respond to. To capture careless responses, in line with best practice guidelines (i.e. Meade & Craig, 2012), participants will be excluded if they respond "no" to the following question: In your honest opinion, should we use your data in our analyses in this study?. No such exclusion criteria was included in the original study but excluded 22 participants in this replication. Everything of this project, including the stimuli, stimuli development documentation, materials, analysis code, summary of results and data have been made publicly available through the OSF project page. Excluding these four differences to the original study, the design is a close replication to that of the original study.
387	7edmy	The replication study deviated from the original study in one notable way. In the original study, US residents were recruited via Amazon Mechanical Turk (MTurk; MTurk workers were required to have an approval rate of 95% of higher and were paid 20 cents for their participation). In the present study US residents via Prolific. Compared to MTurk, participants on Prolific tend to receive higher payments and Prolific advises to pay £12/hour (ca. $15.60; all prices on Prolific are displayed in £). In line with this suggestion, participants were paid £0.60 (ca. $0.80) given that study completion should take approximately 3 minutes. The authors of the original study were kind enough to provide a Qualtrics file of their study, which was also used for the present replication study. Thus, the original design was reproduced exactly.
579	gu26j	Differences From the Original Study. The present study differed from the original study in several ways. The original study used a pen-and-paper survey. The replication study was conducted online. The original study asked participants to report connectedness to their future self by making a mark on a line. A connectedness score was calculated by measuring the distance from the leftmost end of the line to the mark. In the replication study, participants moved a slider (initially displayed as "50") to select a number between 0 and 100. The original study recruited adults who were approached on the University of Chicago campus and at a nearby museum. In the replication study, participants were recruited from MTurk. The original sample was a mix of students and nonstudents. Because the sample was restricted to participants who were 26 years of age or older, I expect that most will not be students. In the original study, participants received a candy bar in exchange for completing the survey. In the replication, participants were paid $0.60 USD. It is unknown whether the original study was double-blind or not. The replication is a double-blind study (prior to analysis). The original study made no effort to identify and exclude inattentive participants. The replication study excluded participants who failed the included attention checks.
7945	k79vj	Deviations from the Original Study 1. The original study relied on data from the fourth wave of the Indonesian Family Life Survey (IFLS), while the replication study uses the fifth wave. 2. The distance to the health center was used as an explanatory variable by Kim & Radoias in the main specification. However, in the replication sample, the variable contains serious limitations. For example, this was only asked of respondents who had visited a medical provider in the last four weeks. Further, only respondents who knew the distance have a value for this variable. Due to the issues with the distance variable, two probit regressions are estimated. The replication study prioritizes the probit regression without the distance variable as the final test for the focal hypothesis. The reason is to reduce the limitations arising from the measurement of the distance variable. However, as Table R.1 shows, the distance variable does not affect the main results of the replication analysis: among the sample of respondents in poor general health who were found to be hypertensive during a screening, the probability of being undiagnosed decreases with education. 7
z516	kbxnr	Deviations from the Original Study Boehm et al. (2015) used data from the Australian Longitudinal Study of Ageing (ALSA; https://www.maelstrom-research.org/mica/individual-study/alsa). For this replication, a 'sister' 3 study, the English Longitudinal Study of Aging (ELSA; Banks et al., 2019; https://www.elsa- project.ac.uk) was selected. - The original study coded education as "less than a high school diploma, high school diploma or equivalent, some college or vocational training, or bachelor's degree or more." The replication study coded education in the UK standards as "no qualification, NVQ1/CSE, NVQ2/GCE O level, NVQ3/GCE A level, higher ed below degree, or NVQ4/NVQ5/degree". - The original study included depressive symptoms at wave 1; the replication study did not have this information available at wave 1, so wave 2 was used instead. The original study used the five-item measure with a range from 0-100 points. The replication study used eight dichotomous items with the aggregated range from 0-8 points. - The original study had 9 waves. The replication study included 5 waves and mortality status relating to 6 waves. - The original study used a one-item measure for life satisfaction. The replication study used a 5-item measure and mean score was computed. The replication study did not have life satisfaction data at wave 1.
8y1	w52zk	Deviations from the Original Study In the original study, participants were "enrolled in one psychology and one sociology class at a public university in Hong Kong [who] were approached and invited to participate in the ... study following their respective classes" (Ku & Zaroff, 2014, p. 474). In my replication, participants were enrolled in several undergraduate business courses and were invited to participate in the study through an invitation letter posted to their student information systems by their respective professors. I have converted the Aspirations Index items into the first-person language, which is also the preferred language in the established version of Aspirations Index, to better adapt them to the Turkish context. Lastly, I administered additional measures for an unrelated study but they did not interfere with this replication effort, as they were administered after the completion of the replication measures.
g7g1	y8rgt	Deviations from original study. Data collection was conducted via online sampling: this was not reputed to be a major concern as increasing literature confirms behavioral research conducted online is comparable to lab/in person testing. Strickland and Victor (2020) is a specific example of this for qualitative data. The replication study confirmed this assumption. The online research method rendered sufficiently and all participants provided detailed accounts on the recall task. An average of 95.5 words were written per participant. A homogeneous sample consisting of subgroups based on nationalities was used in the replication study. Moreover, the replication maintained set parameters (number of participants per nationality, age, educational level, fluency in English) for population sampling as this was considered more robust for statistical analysis. The original study recruited the following proportion of participant samples by ethnicity: 80 European American University students, 98 Asian American University students and 60 Hispanic. 54% of the total sample was male and the University students' average age was 20.8). The Hispanic group was composed of working older adults (average age: 37.1). According to feedback provided by original author during the preregistration process, in his study it became necessary to specifically recruit older Hispanics (only speaking Spanish) in order to obtain a sufficient number of participants scoring high in horizontal collectivism (HC), as this orientation would be more prevalent in this ethnic group. Sociodemographic items: the replication collects more sociodemographic information compared to original study. 13 A coding scheme had to be specifically tailored given limited detail about the coding strategy used for the original study.

21487	yd9bg	Deviations from Original Study - As described in the previous section (other covariates), some covariates were not included in the replication regression model confounders as they were missing from the replication dataset (nationality and education). - In the original study, authors do not specify which distress score was taken into consideration for the regression model described in Table 2 of the manuscript (negative, positive or depression). All three scores were therefore utilized in the replication, providing three regression analysis outputs.
5g9	cdqks	Deviations from the Original Study We planned to run this study using the same methods, materials, procedures, and analyses as the original study as much as possible. However, there are a few differences: - Due to the COVID-19 pandemic we have changed the data collection site in consultation with the SCORE team. Instead of collecting the data via a Lab at our university we have gathered the data via an online survey. - As the original materials are not available, we have developed the quiz at the beginning of the study ourselves. This quiz might differ to some extent with the original study. - As the original materials are not available, we have developed the tests of the study ourselves. These tests might differ to some extent with the original study. - We have awarded standard fees from the survey company instead of course credit. At our department it is not possible to award course credit. However, there can be an effect on the findings here, although it is unclear in which direction it would go, as there are no 'good' or 'bad' answers for the focal hypothesis, contrary to performance games (Bowen & Kensinger, 2017). - The framing "I chose to allocate [A] points, I did so because" slightly differs from the original study, where the framing is "'I chose to allocate the number of points I did because:". We chose to do this as this might help participants to recall the number of points they allocated to the shared account. It is noted that it may serve to increase the detected effect size by priming participants to think more carefully about the exact quantity.
9kgg	n4kby	Deviation from the original study The main deviation from the original study will be the language of the study (Italian instead of English). Additionally, the current coronavirus epidemic in Lombardy hampered the possibility to perform the study in the laboratory. The data collection was thus performed online using Qualtrics.
657	snrcv	Deviations from Original Study: Notably, this effort differed in methodology from the original study in that (a) data were collected from an online survey completed by respondents on the Prolific platform rather than an in-person job fair, (b) data were collected from individuals residing both within and outside of the United States and internationally, whereas the original effort only collected data from individuals present at the U.S.-based job fair, and (c) participants were compensated $4.00 upon completion of the survey in lieu of receiving a keychain. Further, as the original phrasing of the mature job seeker desire items was not available, we used the item descriptors provided in the paper to generate the items used for this replication effort. In addition, we excluded three job seeker preference items from the replication effort, as they did not pertain to the focal hypothesis and included Morgeson & Humphrey's (2006) Work Design Questionnaire to assess additional work characteristic preferences on an exploratory basis.
46z8	tywer	While we did not deviate from our pre-registration, we used mTurk to distribute our study viaSurvey Monkey. Additionally, the original study had 106 participants whereas this replicationhad 262 participants. Similar to Netemeyer et al., we reverse-coded scores on the CurrentMoney Management Stress (CMMS) scale "so that both scales were coded with higher numbers being more positive" (Netemeyer et al., 2018, p. 82). In our pre-registration, we did indicate that reverse coding on the CMMS would occur.
yzm6	7xwyg	The replication differs from the original study in the following ways: 1. The original study had the person-month as the unit of analysis, and used event history analysis; the replication has persons as the units of analysis and uses simple logistic regression. 2. The data set used in the original study had information about when was the last time the respondent had sex with their spouse, and defined sexual inactivity as not having sex with the spouse in the past 12 months. Instead, the replication data set asks about the number of sexual partners the respondent has had in the past year, which was used to derive an indicator of whether the respondent had sex in the past year. By limiting the sample to only respondents who have had 0 or 1 sexual partners in the past year, this 0/1 variable was considered a measure of sexual inactivity with one's spouse, assuming that for respondents who had one sexual partner in the last year that partner was the spouse. 3. The replication does not include "cohabitation", which was a control variable in the original analysis, but is not available in the replication data set.
89z7	te7z6	Deviations From the Original Study: In response to COVID-19, we utilized online rather than in-person data collection. As we stated in our pre-registration: We believe that online data collection is reasonable in this case. There are many previous studies that have shown that online data collection is not worse than in-person data collection (e.g., Germine et al., 2012; Buhrmester, Kwang, & Gosling, 2011; Peer et al., 2017). Furthermore, many studies in decision making, including several by the lead author of the to-be-replicated study, frequently use online data collection for studies on similar topics (Rich & Gureckis, 2017; Lieder, Griffiths, & Hsu, 2017; Dubey et al., 2018; Lieder et al. 2018; Bourgin et al., 2019; Hardy & Griffiths, 2019). This has become such a standard practice in the field (c.f., Stewart, Chandler, & Paolacci, 2017) that these sample references are only a tiny subset of the relevant work.
61k8	ht4d5	Differences from Original Study Although most replications necessarily take place at a different time and place than the original study, it is especially worthwhile to emphasize that data collection for the replication began just before large portions of the United States began to implement various types of stay-at-home orders as a result of the the SARS-CoV-2 pandemic. Although we can't know precisely how this might have impacted participants' responses (although this might be an interesting question for future research), given the politicization of pandemic-related response strategies in the United States, the data reported here should be interpreted with caution.
k17	uwqs5	Deviations from the Original Study: In the original study, data was collected using paper and pencil questionnaires. Our study used an online questionnaire.
8g91	j24ts	The study did not substantially deviate from the original one, except for the fact that all study materials was translated into Polish. Moreover, at the end of the study additional six-item questionnaire of quest for significance (Molinario et al., 2022) was applied. The questionnaire was administered after completing the main study (before debriefing), therefore, it could not affect the results in any way.
y071	69zjt	Deviations from the original study: 1. Given that the initial study also utilized MTurk, we added an additional exclusion criterion, namely whether participants have already participated in this survey before ("To the best of your knowledge, have you ever previously participated in this particular study (i.e., "Intellectual performance in the presence of a co-actor") or a nearly identical one before?"). Participants were excluded if they answered "Yes." 2. For the exclusion criteria that participants did not believe that they a) would be completing another three rounds of the anagram task in the second part of the study and b) that they would be matched with another participant, we had two answer options ("Yes" or "No") as compared with the original study that also had the third option "I had some doubts." 3. It seems that in the original study both other sham co-participants were described as P81 (see p. 404). We deemed that it is illogical that two alleged different persons would receive the same code and changed one of these to P33. 4. As noted previously, we analyzed the data using Mplus 8.4 rather than STATA (given that we are not trained in STATA). The script and code were derived from Sommet, N. and Morselli, D. (2017). Keep calm and learn multilevel logistic modeling: A simplified three-step procedure using Stata, R, Mplus, and SPSS. International Review of Social Psychology, 30(1), 203–218, DOI: https://doi.org/10.5334/irsp.90)
23wwg	srk8q	Deviations from the original study: 1. The original study uses the World Values Surveys (WVS), rounds 3 to 5. The data used in the replication is from the same source as the original study but using rounds 6 to 7. 2. The "Formal Education" variable is obtained in the original study using a nine-point scale. However, the same scale is not available in WVS Wave 7. Instead, "Formal Education" is constructed using a three-point scale. 3. In the original study, the democratic freedom scores are weighted down for the presence of human rights repression. However, the human rights repression is obtained from Cingranelli and Richards Human Rights Project, and the latest year that is included is 2011: http://www.humanrightsdata.com/p/faq.html Because WVS Wave 7 started in mid-2017, there is no available data to adjust down the democratic freedom scores in the replication.

19gz	xg6q5	Deviations from the Original Study (Screening) In the original paper, the authors exclude illiquid stocks, based on the following criteria. “We exclude from our sample stocks with missing volume greater than 1% of the number of trading days in the study period.” (page 478) There were 756 trading days in their study period (Jan 2007 to Dec 2009), and therefore, it implies that the authors remove stocks if they do not have more than 749 trading days. This screening process excluded about half of NASDAQ stocks in their sample and only 1,743 stocks satisfied this condition. We do not have any problem doing the same screening process applied in the original paper. However, we do not think it is a sensible thing to do to our extended sample. For example, we collected NASDAQ stock data over the 2010-2019 period, and there are 2,510 trading days. If we apply the same screening method (i.e., remove all the stocks that do not have more than 2485 daily trading records over the 10 year periods), only 949 stocks (2.34M observations) satisfy this condition. This screening will essentially eliminate delisted stocks and newly listed stocks over the last ten years. This may lead to a selection bias problem. As a result, in addition to the original screening process our team decided to consider a bit weaker screening process – exclude stocks for a year if the stock has missing volume greater than 1% of the number of trading days in that year. Under this screening process, we find that there are 3,804 stocks (4.85M observations) in our sample over the 2010-2019 period. We created two datasets – one dataset based on the screening process suggested in the original paper, and the other based on a weaker screening that we suggest above. (1) “Final_Data_OriginalScreening.csv” is a replication dataset created based on the screening process suggested in the original paper, and (2) “Final_Data_AnnualScreening.csv” is a replication dataset created based on our suggestions described above. We use the “Final_Data_AnnualScreening.csv” dataset in our replication analysis. Deviations from the Original Study (Data)The replication data are from 2010-2019 while the original uses data from 2007-2009.All the variables from the same source except the two variables – (1) ‘quoted spread’ and(2) ‘price volatility’.(1) Quoted Spread:The original authors collect the national best bid and ask prices from TAQ (Trade and Quote) database. However, we construct quoted spread from daily bid and ask prices provided from CRSP (Center for Research in Security Prices).The team believes this deviation is justifiable as the same author (Chung) documents that quoted spread constructed from CRSP bid and ask price data is a good approximation for quoted spread constructed from TAQ data. They find that the correlation between the quoted spread constructed from CRSP data and TAQ data are no less than 0.92 for 1993-2009 period [Table 1, Panel B]. Chung, K. H., & Zhang, H. (2014). A simple approximation of intraday spreads using daily data. Journal of Financial Markets, 17, 94-120.(2) Price Volatility:The original authors construct daily price volatility from five-minute quote mid-point stock returns in TAQ data. However, we construct the same variable from daily hi and low prices available from CRSP, based on Parkinson (1980)’s earlier (seminar) work. Constructing daily volatility from daily high and low prices is a common practice in Finance. Parkinson’s work has been cited 1898 times. Parkinson, M. (1980). The Extreme Value method for Estimating the Variance of the Rate of Return. The Journal of Business, 53, 61-65.
6m492	wygup	While the interval between two data collection points is comparable for both datasets, it does differ by a few months. This could potentially impact the strength of the correlations between the waves, making the effect more difficult to detect. The advantage of the replication dataset, however, is that it contains more than twice the observations, increasing the power to detect smaller effects. Tthere are also two main differences between the samples. One is that the original sample used students as opposed to the general population. However, as the replication dataset has a larger and a higher quality sample, it should offer a good-faith replication of the original study. Second is that the replication dataset was collected during a pandemic whereas the original paper was published before this period of time. However, this can be beneficial to the test of the hypothesis as authoritarianism and depression levels were more likely to fluctuate during the pandemic (i.e., external threats such as the threat of pandemic can increase depression levels and authoritarianism levels among individuals to a different extent)., Thus, the context of the pandemic and its impact on both depression and authoritarianism can provideing a robust test of whether a change in authoritarianism would predict a change in depression.
6m34m	2jmtf	The original study applied data only from China Statistics Yearbook On High Technology Industry from 2002 to 2005, which contains data from 1995 to 2004. The Yearbook of 2019 is published, but the dataset is only available until 2017. Every Yearbook has 6 years of data, for example, the Yearbook in 2011 contains data from 2000, 2005, and 2007 to 2010. For the replication, Yearbook 2002, 2007, 2011, 2015, 2017 were applied to cover the data from 1995 to 2016.FixedAssets is not available after 2010, which leads one of the control variables CAPINT is not available. Output is also not available after 2011, which is also the ingredient of CAPINT. Therefore, the analysis can be done either without CAPINT and include every year,or with CAPINT and only apply data before 2010. The former one is now applied in the analysis. That is the only deviation from the original study.
m707	qk34m	Deviations from the Original Study The original study was conducted in person, whereas the replication study was conducted online. The original study manipulated target race, but the replication used only a Black target. In the original study, participants received course credit for their time; in the replication study, students were paid with $5 and $10 Amazon gift cards to participate in the first and second waves of the study, respectively. The original study plotted the effects of communication positivity (subjective ratings) at 1 SD above and below the mean of IMS. In the replication, however, 1 SD above the mean of IMS was beyond the maximum observed value for IMS; accordingly, the effects of communication positivity were plotted at 1 SD below the mean and at the maximum value of IMS.
2y486	52nkw	In the original study, the researchers found that mortality salience impacted utilitarian judgments of moral dilemmas where “doing the most good” required bringing harm to an innocent bystander, which is known as “instrumental harm” in utilitarianism. However, the concept of utilitarianism consists of two primary elements: instrumental harm, as described above, and impartial beneficence (Kahane et al., 2018). Impartial beneficence is the impartial concern for the well-being of all peoples, regardless of relationship or proximity (e.g., one must be willing to do the greatest good at the expense of close friendships). In this study, I aim to test whether or not mortality salience impacts people’s utilitarian judgments in dilemmas where people have to decide to do the greatest good at the expense of helping a family member in need.
6m396	atwck	The study will speak to the original claim 4 by generalizing the results to moral dilemmas with non-physical harm (as opposed to just physical harm), which taps into the other important dimension of utilitarian judgments: impartial beneficence.
8m17	qu793	Deviations from the Original Study The study sample was very similar to the original, with age range from 19 to 62. Eighteen of the 20 participants were run in the lab, as in the original study. However, two participants were run online through chrome remote access.
675m9	u58za	Deviations from the original study: 1. The original study uses the Comparative Study of Electoral Systems (CSES), Module 3. The data used in the replication is from the same source as the original study but using Modules 4 and 5. Each module corresponds to a different survey wave: • CSES Module 3 data was collected from 2006 through 2011. • CSES Module 4 data was collected from 2011 through 2016. • CSES Module 5 data was collected from 2016 through 2021. 2. For each considered country, the most recent data is used. The original study obtains the measure of valence using party affect, leader affect, and perceived party competence. However, perceived party competence is not available in CSES Modules 4 or 5. Therefore, in the replication, valence is obtained based on party affect and leader affect.

z106	rak92	Deviations from original study. There were several deviations in the method from the original study, that were all highlighted in the preregistration. They were: - Our sample are UK based respondents that were collected via Prolific. The original study had a sample of 195 university students (age 18-29) in Australia. Our sample thus deviates in sample composition (age range / education level) and with regard to sample location. - We made some adjustments to the newsletter, to adapt it to the UK instead of Australia. o First, for the neutral stories we changed a) a story from a for-profit utilities provider in Australia, to another story of one in the UK and b) a story about an organisation for care for the elderly to an organisation with a similar name and purpose that is operational in the UK. o In the description of KFC and Subway, we adapted their local origin stories to reflect their presence in the UK (notably the year of entry in the market, the number of stores, and number of employees in the UK). o The most important changes is that Red Cross Australia and the Leukaemia Foundation were too location specific, so we changed these to the British Red Cross and Leukaemia UK, which are similar but based in the UK. For Leukaemia UK we could keep the story the same, but for the Red Cross we had to adjust the situation a bit, as in Australia the Red Cross takes care of blood donations (as mentioned in the newsletter created for the original article), but in the UK the National Health Service (a government agency) does this. We therefore chose to describe the emergency support the British Red Cross provides, to maintain the same organization and the same local focus in its described activities. Finally, the events that were sponsored by either KFC or Subway were changed to take place in the UK. - Our final preregistered SEM model differed slightly in the degrees of freedom compared to the model reported by the original authors. All specifications are explained in detail in the preregistration plan and can be found in the R code for the analysis.
67g8	8npm5	Deviations from the original study: There are a number of deviations from the original study documented in the preregistration (https://osf.io/8j3gk). These are all considered minor, with the exception of one. Srinivasan and Carey (2010) collected their data in a laboratory. We collected ours online. This deviation was necessary due to the COVID-19 ‘lockdown’ in the first half of 2020.
7976	ad9jm	As detailed in our preregistration, our only significant deviation from the original study was that in our study, the person collecting the data was blind-folded during the estimation process so that the experimenter could not see which arm the participant was estimating the length of. Since the researcher could not be blind to the handedness of the participant, this ensured they were nonetheless blinded to the critical handedness x measured-arm interaction.
4zy8	gaqbx	Deviations from the Original Study The original study was on paper and the replication study is online. This online study requests but does not require a response for each question. In the original study, participants were “enrolled in one psychology and one sociology class at a public university in Hong Kong [who] were approached and invited to participate in the … study following their respective classes” (Ku & Zaroff, 2014, p. 474). In our replication, participants were enrolled in undergraduate psychology courses and signed up to participate through our department’s online research management site. We made two small wording changes to the main survey for better fluency in English: Item 15: "You will work for the betterment of the society." to "You will work for the betterment of society." Page with Willingness to Pay measures: instructions from "Do you agree with the following statements?" to "How much do you agree with the following statements?" In our questions regarding subject demographics, we changed the word “sex” to “gender”, changed the gender options to reflect the culture at our college (male, female, other, prefer not to answer), and made the school options relevant to the schools of our college (Business, Communications, Health and Human Performance, Humanities and Sciences, Music). Furthermore, since our student population is diverse in race/ethnicity, we asked students to identify their race/ethnicity so that we might better describe our sample in our report. It should be noted that the differences between our mostly white, but diverse, sample from the United States, and the original study’s homogenous Chinese sample, presented a deviation from the original study that could potentially have a bearing on the results of the study. The method of compensation differed between the original and replication study. In the original study, students “participated out of their good will” per the corresponding author, and were provided a small packet of confectionary as a token of appreciation. In the present study, students received extra credit in their psychology course for their participation.
16yz	6aefs	Deviations from the Original Study There are no known deviations from the original study.
zz26	udbnf	Deviations from the Original Study The major deviation between the original study and the replication is the national context, Germany vs. USA. In addition, the original study did not specify the ethnicity of the participants, whereas in the present study we explicitly sought to recruit a sample that consisted of White, Black, and Latino Americans. The original findings were taken to hold for adults in general, with the authors suggesting generalization across Germany and the U.S., so these deviations are not expected to impact the ability to replicate. Finally, we added a number of scales not used in the original study, as detailed in the Measures section of the preregistration, although the target scales were always presented at the beginning of the survey, and therefore these additions should have no impact of the results.
658g7	dkcu6	*the original study puts people who were at least 21 in 1991 in this category. This replication only has age data in brackets therefore it is impossible to single out only people who were 21 in 1991. Instead age people within age brackets 35-64 and 65 and up (in 2014) were included in this category. This makes the youngest participants 12 in 1991 which can be reasonably classified as growing up in the Soviet Union).
546	f5esc	Deviations from the Original Study For the replication study, the 2011 survey rather than the 1998 ECLS-K survey was used so this is an entirely new dataset with entirely new subjects. There are some differences between the 1998 and the 2011 questionnaire. The ECLS-K used three subsets of the preLAS2000: "Simon Says", "Art Show" and "Let's Tell Stories" to evaluate the child's natural speech. These three subtests were referred to as the Oral Language Development Scale (OLDS). In the new ECLS-K: 2011 data, only two of the preLAS subtests, “Simon Says” and “Art Show” were used to evaluate language skill. The "PRELAS SIMON SAYS SCORE" and "PRELAS ART SHOW SCORE" range from 0 to 10. Due to the above deviation, different from the original paper that uses 37 points in the Oral Language Development scale to define ELL, the new survey has its new definition: “English language learner (ELL): A student whose native language is one other than English and whose skills in listening, speaking, reading, or writing English are such that he or she has difficulty understanding school instruction in English”. (https://nces.ed.gov/ecls/pdf/kindergarten2011/Fall_K_Classroom_Teacher_Teacher_Level.pdf) Variables that are similar to the ones picked from the original study have been chosen so that a faithful replication can be made. In the original study, students self-reported internalizing and externalizing social emotional problems were analysed. However, as this was not available, the teacher-reported externalising and internalising problems have been chosen alongside student self-reported peer victimization and anxiety. These new variables have been chosen based on how well they resemble the original study and not if they are likely to give confirmatory result. The academic outcomes variables are identical to the variables used in the original study.
6zzo6	dnqrv	Deviations From the Original Study: We used Prolific, an online labor market designed specifically for online research, rather than Amazon Mechanical Turk (MTurk). Pay rates are higher on Prolific, and higher due to inflation, so we adjusted the pay accordingly. We offered $2.50 for completing the task, and up to $2.00 in bonus pay. The only other deviation from the original study is the choice of analysis model. To conform with SCORE’s requirements, we used frequentist inference instead of the Bayesian inference used by the original study. While there are a variety of theoretical reasons to prefer Bayesian inference even in relatively simple statistical models like this one (Kruschke, 2013), in practice frequentist inference provides a similar bottom-line conclusion in most cases. To demonstrate this, we reanalyzed the data from the original study using our SCORE-compliant analysis plan. We found that the 1D scores for learners in the contingent information group were significantly higher than the 1D scores of learners in the full information group, $t(93) = 2.91$, $p = 0.0045$, 95% CI = 0.027 to 0.145. We reach the same conclusion, and the 95% frequentist confidence interval matches closely the 95% Bayesian credible interval.
z189	2f35a	We did not depart from the original study in any meaningful way. We used the original materials, the original procedure, and sampled from the same participant population as the original study.
3z5z	dw4mf	Deviations from the Original Study The original study used real state GDP chained to 2007 dollars, but we used it chained to 2012 dollars. See the next section for one additional deviation in the estimation approach.

z4z9	jr239	Deviations from the Original Study 1. Although the data sources are the same (the IDEA Data and the Digest of Education Statistics), the original study relied on data from 2014, while the replication study uses data from 2017.
5z36	unj3f	Deviations from the Original Study In the replication data sets, there were only two indicators of the valence dependent variable rather than the three used in the original study. Therefore, unlike the original study, it was not possible to create a principal components score for valence ratings with only two indicators, and an average of respondents' two available valence ratings were used as the primary dependent variable instead. Also, for one of the political parties rated, there were two separate leaders who contributed to respondents' valence ratings, one who had just finished serving as the head of the party and one who was running as the new head of the party in the current election. Therefore, one party had three scores contributing to the average valence ratings, whereas the other parties had only two.
6zz3k	5869t	The present replication attempt expands on the original study and uses almost 24 years (instead of 13 years) of monthly-level reports from the NLSY97 dataset. More detailed comparisons, including all the known deviations from the original study, are discussed in the preregistration (https://osf.io/tkd3q).
2g7ky	8mqzu	This study follows Weidmann and Callen (2013) to investigate the relationship between violence and election fraud. While the original study examined cross-sectional observational data from the 2009 Afghan presidential election, the replication study uses data from the 2014 Afghan presidential election.One critical difference between the 2009 and 2014 data is that no candidate won a majority of votes in the 2014 election and, therefore, a runoff election was held afterwards between the top two candidates. The 2014 dataset is thus consisted of the pooled voting results for both elections. Moreover, in 2009, the votes from a random 10% sampling of boxes (N=342) were recounted due to allegations of fraud. The authors used the results to estimate the share of fraudulent boxes in each district. By contrast, in 2014, the disputed run-off election was recounted in full – votes from all boxes were recounted. Therefore, the district-level share of fraudulent boxes is observational data instead of statistical estimates.In addition, the 2014 recount did not directly identify any box as "fraudulent." Instead, eachbox was marked as "Validated," "Recounted," or "Invalidated" in the post-auditing records. Note that, due to the ongoing political crisis in Afghanistan, all official websites hosting election data and relevant explanations are not accessible (as of April 2022). Against this backdrop, how to operationalize "fraud" is a challenge of the present study. Eventually, five variables were created as parallel measures of election fraud, and all five of them would be tested in the replication. (Please see the preregistration for a detailed discussion.)Lastly, a few measurements in this study differs from the ones used in the original study. 1)the geographical variables "elevation" and "distance from Kabul" have been rescaled to a larger unit. Please see "General Discussion" for more details. 2) robust standard errors estimated in the main models are clustered at the province level rather than the regional command level. Nonetheless, additional models with a clustering at the regional command level have also been estimated and reported in "Additional Analyses." Please see "General Discussion" for more details. 3) both the present study and the Phase 1 study used the measure of "Percentage of polling centers closed", instead of 'Number of closed polling stations.' As explained by the researchers during Phase 1, this is because data on the number of planned polling stations is unreliably measured. Moreover, normalizing the measure as percentage can adjusts for population differences across districts. Despite these differences, the present study does not deviate from the original study. The 2014 dataset features the same variables as those used in the original study, except that they reflect situations in 2014 and some variables are measured in a slightly different way. The unit of analysis for both studies are "district". In terms of statistical analysis, the present study replicates the same logit and OLS regression models of the original research and no deviations occurred.
6z3o2	cnbt5	The data source for replication is the same as in the original study, however, a different wave will be utilized. The PISA study 2012 has been chosen for replication, because it is the most recent wave with a focus on mathematical abilities. PISA 2003 was administered in 41 countries (N = 276,165)and PISA 2012 was administered in 65 countries (N = 480,174). However, the final analysis sample for the replication attempt will be much smaller due to the fact that in PISA 2012 certain items were not administered in all countries(see below). The final dataset includes all variables to conduct the analyses of the marked bushel claims. The main analysis is a weighted multi-level model considering the individual, the school and the country level. The selected variables in the final dataset are the same ones that were used in the original study. We will use all available countries in the PISA 2012 data for the replication analysis, even if some were not included in the prior original analysis. The reason for this is that the authors of the original study also do not limit their interpretation to certain countries, but use all data available. Detailed information on which countries were included in PISA 2003 and 2012 can be found in the Technical report of PISA 2003andPISA 2012(also uploaded to the OSF project).
g9mm	gsw9a	Deviations from original study The sample in the original study were recruited form a German university and received partial course credit for participation. The sample for the replication were recruited through online crowdsourcing platform, Prolific, and were remunerated with £6 or around $10 CAD. This divergence from the protocol meant that it was not a direct replication in the same population but the hypothesised effects should apply to a general population. In the original paper, the methods section described how participants were asked to "name their most important interpersonal concern". The materials in the Kappes et al. (2012) study were in German and we chose to use the term "goal" as we felt that "goal" was an appropriate substitute for "concern". The word "concern" in English also means something that causes worry or anxiety in addition to meaning a matter of importance. Therefore, the word "concern" could be misconstrued by participants. For example, use of the word "concern" may have elicited a statement which does not include a desired direction or aim (e.g. "my boyfriend and I are having problems"). The word "goal" is clearly understood in English to represent a direction or aim for an outcome and so was unlikely to be misunderstood. This replication deviated from Kappes et al. in the number of trials that participants completed. The total number of trials may had been reported incorrectly in the original paper. Kappes et al. (2012) reported that in addition to the 16 critical trials, "Thirty-six filler trials containing neutral words as primes and as targets (e.g., umbrella, noon) and 48 non-word trials were included. Thus, the complete lexical decision task contained 96 trials. Half of these trials were real word trials, of which one-fourth were critical trials." However, this sums to 100 test trials, 52 of which were real word trials. Therefore, the replication had 34 neutral word filler trials, instead of 36, and 50 non-word trials as opposed to 48. Thus, participants were presented with 100 trials, half of which were real-word and half were non-word trials. In addition, there was a block of 10 practice trials comprised of non-word prime-target pairs and unrelated word prime-target pairs in the replication.
10g2	2cpf8	Deviations from the Original Study The original paper uses the U.S. data from the 1996 Panel Survey of Income and Program Participation (SIPP). The SIPP is a household-based survey designed as a continuous series of national panels. Each panel features a nationally representative sample interviewed over a multi-year period lasting approximately four years. The 1996 SIPP tracks all members of a household and splinter households for up to 51 months over 1996-2000. Following the original paper, the latest 2014 panel of SIPP is used to construct the replication dataset. It includes four waves, covering January 2013-December 2016. Other deviations are minor: the interaction variables are identical except there is a slight difference in the firm size variables. The firm size categories for the original study are 1-24, 25-99, and 100+. They are 1-25, 26-100, and larger than 100 for the replication. The original study reported 223 industry categories. The replication data has 258 unique industry categories The original study used the filter Wage <=12,500. For the replication this filter is adjusted for 2014 using the CPI. This produced a wage filter of Wage <= 18,858
yzz0	rg5vt	Deviations from the Original Study (1) Four different variables (family income decile, raw family income, family income brackets, and raw personal income) were used to create income decile, whereas only two variables (i.e., raw family income and family income brackets) were used in the original study. (2) Three control variables are different from the control variables used in the original study: (i) church attendance rather than self-assessed religiosity measured religiosity; (ii) party affiliation rather than left-right self-placement measured left-right orientation; (iii) Gini coefficients were gathered from the WIID rather than from Solt (2016); (3) Two control variables used in the original study were missing from the ISSP data, namely, fixed-term employment and establishment size. We replaced these variables by "public sector employment" (that the authors wished to control but could not) to provide relevant alternative information regarding respondents' type of employment/establishment.
ykk0	uvgp5	Deviations from the Original Study There were no deviations from the original study.

m7m3	5qet4	Deviations from the Original Study In deviation from the original study, only measures of environmental attitude were included and analyzed in the meta-analysis. In relation to this deviation, a univariate meta-analysis was used instead of multivariate one.
21k52	r6edv	For the replication of the biyearly sample form the year 2012 to 2018 have been chosen compared to the original sample years 2006, 2008, and 2010.
g241	j3u2s	Deviations from the Original Study There are no deviations for the reproduction data The replication data extends the dataset (from 2002) to 2013 but data for France are excluded as government ideology for France is no longer available in the updated dataset.
93k7	h7jfv	Deviations from the Original Study To conduct the Latent Class Analysis, the original studies uses MPlus, while this study uses the Lclass extension to the GSEM package in Stata 16. Otherwise, this study contains results from the same GSS years used in the original paper (2006, 2008, 2010), years not used by the original paper (2012, 2014, 2016, 2018), and a combined analysis with all eligible years from the GSS (2006, 2008, 2010, 2012, 2014, 2016, 2018). The central replication point reflects the years not used in the original study. I tested the analytic sample my code produces using the same years as the original study. I found that my sample contains 90 more respondents than the original study. It is unclear how the original authors decided to remove those observations. I do not believe this difference changes the outcome of the focal hypothesis test.
6m33m	qfh2r	Deviations from the original study: 1. The original study relied on data from the 2008, 2010, and 2012 AmericasBarometer national surveys of Argentina and Mexico. The data used in the replication are from the same source, and for the same countries, but from the 2014, 2016/17, and 2018/19 surveys. “These data were supplied by the Latin American Public Opinion Project at Vanderbilt University, which takes no responsibility for any interpretation of the data.” (see data’s terms and conditions at http://datasets.americasbarometer.org/database/agreement.html). 2. In the original study, the sociotropic economic evaluation is based on the question: “How would you describe the country’s economic situation? Would you say that it is very good, good, neither good nor bad, bad or very bad” (page 1 Appendix) However, this question is only asked for Argentina in 2014. In the replication, the sociotropic economic evaluation is based on the related question: “Do you think that the country’s current economic situation is better than, the same as or worse than it was 12 months ago?” 3. In the original study, Model 1 in Table 1 (page 1335) includes the support for the notion that the state should be the primary source of employment (“Government should provide jobs”) as a control variable. However, the question is not available on the 2014, 2016/17, and 2018/19 surveys. Therefore, it is excluded as a control variable in the replication. The implication of not including the variable is that the model will not capture how citizens' preference for the state as a primary provider of employment differs in supporting the incumbent party in dominant-party systems. However, this variable was non-significant at conventional levels in the original study (see Table 1, page 1335). Therefore, it should not affect the replication outcome. 1 Moreover, Model 1 in Table 1 (page 1335) includes a pocketbook economic evaluation based on the question: “How would you describe your overall economic situation? Would you say that it is very good, good, neither good nor bad, bad or very bad?” (page 1 - Appendix). However, the question was only asked for Argentina 2014. In the replication, the pocketbook economic evaluation is based on the related question: “Do you think that your economic situation is better than, the same as, or worse than it was 12 months ago?” 4. In the original study, Models 1 and 3 in Table 2 (page 1339) include the “Government security evaluation” as a control variable. However, the question is not available on the 2016/17 and 2018/19 surveys. Therefore, it is excluded as a control variable in the replication of Trace number: 4 / Claim ID: l8rz5d and Trace number: 5 / Claim ID: 6rp9on. The implication of not including the variable is that the model will not capture how citizens' evaluations of security provision differ in supporting the incumbent party in dominant-party systems. It should not affect the replication outcome as long as it is not related to the sociotropic evaluation (the authors of the original study do not present results using each variable separately).
2y2g	5fvdy	-The nationality of the sample is different. In the original study, they used a sample of 2,835 German 6th-graders (Arabic-German: n= 105, Chinese-German: n = 110, Polish-German: n = 57, Turkish-German: n = 383, heterogeneous bilingual: n = 284, and monolingual German group: n = 1896). In the replication study, we used a sample of 20,026 students from 14 differente participating countries (Belgium, Bulgaria, Croatia, England, Estonia, France, Greece, Malta, Netherlands, Poland, Portugal, Slovenia, Spain, Sweden). The claim that the original authors make might only apply to German participants, and not to participants from other countries.-In the original study, authors assessed English ability with a different instrument. Specifically, original autors used the Cloze test (consisting of four texts with 91 word completion questions measuring reading proficiency, vocabulary, grammar and spelling simultaneously). In contrast, the proposed replication dataset includes measurements of students listening, reading and writing abilities taken from The European Survey on Language Competences (ESLC) from 2012. In the replication, a composite English ability score between these three dimensions was calculated. This composite score might be biased because it is an average of three dimensions (students' listening, reading and writing abilities). In fact, check up analyses carried out by the replication team found that students that responded to the wirting ability test scored, on average, significantly lower than those who responded to the reading and listening skills. That is, it seems that these three scales might measure different dimensions of the construct “English ability”. In this regard, check up analyses were performed, and the main claim was tested for each category of English ability. -The average age in the replication sample is statistically larger than the mean age of the original sample. The mean age in the original study was 12.71 for the bilingual group (SD= 0.70, 95% CI [12.87, 12.75]) and 12.49 for the monolingual group (SD= 0.49, 95% CI [12.47, 12.51). In contrast, in the replication sample, the mean age for the bilingual group was 15 (SD= 0.89, 95% CI [14.99, 15.01]) and for the monolingual group was also 15 (SD= 0.92, 95% CI [14.97, 15.03]). The 95% confidence intervals show that the mean age of the replication sample was significantly larger than the man age of the original sample. This difference in the mean age could lead to substantial differences in the findings of the original and replication study.-In the original study, authors used as a control variable a measure of cognitive ability. Specifically, they mentioned “As general cognitive abilities might systematically differ across groups, we used a composite score of two subtests of the CFT4-12R: verbal and figural analogies (Heller & Perleth, 2000). This test consists of 25 picture and 20 word tasks subtests and was administered in the fourth grade. TFor cognitive ability, the replication study does not have this measure.” (page 79 in the original paper). There is no such a measure in the replication database to control for. Thus, this is another difference between the original study and the replication study. Not controlling for this variable might lead to differences in the findings from the original and replication study.
288g2	t3py6	Deviations From the Original Study: In addition to the two main ways we are testing generalizability (a less-informative-prior scenario and randomly sampling the numbers for each participant), we also deviate from the original study in a few other ways. These could be considered other minor checks of generalizability. 1. The original study was conducted with undergraduates at a US university. We sampled from an online labor market (Prolific). 2. The original study was conducted by having participants read the questions on a piece of paper and write their responses. We computerized the task. 3. The original study presented 1, 3, and then 10 samples. We add one sample at a time, but only ask for an estimate after the 1st, 3rd, and 10th. 4. Our analysis procedure relies on a set of t-tests rather than an ANOVA. 5. We used a slightly different rule for outlier exclusion, based on recommendations in Leys et al. 2013.
8zz7	xt4j7	Deviations from the Original Study One main deviation is that the original study used survey weights. These weights appear to not be available in the current data set.
1y02	dscfg	Due to research restrictions associated with COVID-19, the replication experiment was conducted online rather than in-person. We therefore had little control of the experimental environment/time/etc., which may have affected participants’ attention, motivation, etc. We did attempt to avoid including data from participants who were not attending to the task by periodically asking if they were paying attention (after blocks 3, 6, 9, and 12) and excluding those participants who reported not paying attention for more than two of the four question prompts. This led us to exclude and replace data from 15 participants, however it is still possible that even the participants who did report that they were trying were still paying relatively little attention compared to the in-person participants in the original study. One hint that this may be the case is that only 30% of participants responded “yes, lots of attention” for all four prompts (although note that 75% responded with either “yes, lots of attention” or “yes, mostly” for all four prompts). Another hint is that our observed d-prime values were not above chance in either condition (random: mean = -0.0045, t = -0.21, p = 0.83; starting-small: mean = 0.055; t = 1.01, p = 0.32). Other minor deviations from the original experiment – in particular, the change from Dutch- to English-speaking participants and some related changes in the specific syllables used in the to-be-learned grammar – seem less likely to relate to the unsuccessful replication.

4142	8hnxg	Deviations from the Original Study It is unclear how the original study dealt with the empty cells in the dataset and, therefore, our replication might deviate from the original study.
kzyz	pex6u	Deviations from the Original Study There are no substantial deviations from the original study.
658y7	skwev	This study differs from the original study in the screening process. The original study uses Oral Language Development Scale (OLDS) to identify English language proficiency. The 2011 cohort that will be used in this replication was screened with a preLAS scale. The only difference between the two scales is that OLDS uses three tasks (Simon Says, Art Show, and Let's Tell Stories) and preLAS uses only two tasks (Simon Says and Art Show). The tests are used to assess English proficiency and both original study and replication uses the test to select only English-proficient children for the final analysis. This will not affect original claims.
m5k3	3cp2d	Deviations from the Original Study While there are no deviations between the pre-registered replication protocol and the actual study to report, there are several differences between the original study and the present replication study that should be noted. First, the present sample was considerably more racially diverse than the original sample. While both samples were majority-White, this was truer for the original sample than for the present one; the original sample was approximately 93% White, whereas the present sample was 60.46% White. The original study's first author identified this as a potential constraint on generalizability prior to the present study's pre-registration. The pre- registration for the present study therefore included an analysis using only data from White participants. Among only these participants, a very small, yet statistically significant, effect was found in the predicted direction in Phase 2. This discrepancy is therefore one plausible explanation for the null effect observed in the primary model. Second, while the original study was longitudinal, the present study is cross-sectional. Relatedly, the original study sampled participants between 14 and 25 years of age and estimated the relationship between sexual orientation and desire for toned and defined muscles across late adolescence and early adulthood, whereas the present study sampled participants only at the higher end of this distribution, between 18 and 25 years of age, and entered age into the models as a covariate. Additional discrepancies include the geographic populations sampled (i.e., the original study was an exclusively American sample, whereas the present study sampled from six Western, English-speaking countries), the way in which height and weight were measured (i.e., the original study measured height and weight using open-ended prompts, whereas the present study asked participants to select their height and weight from drop-down menus capped at 4 feet and 7 feet and 50 pounds and 400 pounds, respectively), and treatment of missing data (i.e., the original study included participants with missing data in their dataset as long as each had at least one measurement for each of the key variables, which was possible due to the study's longitudinal design, whereas missing data were dealt with through listwise deletion in the present study).
8z2g	w9acu	Deviations from the original study The replication data deviates from the original study in several important ways. First, the original study was conducted in a sample of high school students, whereas the replication data come from a sample of college students. As a result, the GPA variable in the replication may be a more noisy measure of academic success given that there is a wider distribution of course difficulties in college, more self-selection into courses, and less reliability in the GPA of a freshman, sophomore, or junior which is based on only 0.5-3 years, than in the high-school GPA based on all four years of high school. Second, the measure of self-regulation used in the paper that the replication data come from is based on just one scale; however, in the original paper, self-regulation was based on four scales. The replication's measure of self-regulation (ASE) may therefore be less reliable or valid than in the original paper, which could attenuate the association with GPA. Third, in the original study, self- discipline was modeled as a latent variable, but the correlations reported in the replication study were not based on a latent variable, which means it is modelled with error which could attenuate its true relationship with both GPA and ASD—potentially making it a less effective control variable than in the original study. Finally, the original paper included gender as an endogenous variable, but the replication data did not include information about students' gender.
38	zhdtn	Deviations from the Original Study The original study and replication samples are very similar. The variables come from the same data sources, albeit the replication study is more recent data. The sample sizes are very similar.
1642	78xj3	Deviations from the original study: A number of deviations were made from the original study: • Participants were paid $0.60 for completing the survey in the original, but in this study we paid participants $1.20 because of updated norms that suggest $0.12/minute is a fair minimum wage for participants on MTurk and the survey was expected to last 10 minutes. • Participant demographics that were not part of the pre-screening criteria were captured at the end (rather than beginning) of the survey to allow participants to address key measures as early in the survey as possible. So we included the following at the end: gender, race, industry worked, occupation, etc. • Participant education, personal income and perceived health were also captured, these were not included in the original study. • Participants were asked to complete a captcha code check and to thoughtfully commit to providing their best answers after completing the informed consent. • Participants in the original study were asked to answer ICS questions twice, referencing first older workers (55–70), and then younger workers (25–40), whereas in our study, the ICS questions were presented only once, referencing younger workers. • We included a new measure at the end of the survey (but before the final sociodemographic questions): Motivation to Continue Working (MCW, from Bal, de Jong, Jansen & Bakker, 2012). This deviation was made to test the additional hypotheses and was not expected to influence survey results as it was completed after other scales, prior to demographics. • The original study surveyed participants in the following order JSS, WICS, AOWS, ICS, and SIC, whereas in our study, participants completed the scales in a randomized order. The order of questions within the scales was not randomized and presented in the same order as the original study. Author correspondence suggested order of scales should not influence outcomes (available here). • Due to the COVID-19 outbreak, many workers were working under alternate circumstances. Thus the following message was included immediately prior to survey measures, "Please Note: We understand that the COVID-19 situation and social distancing guidelines may be causing disruption to your current work routines and interactions with co-workers. In answering these questions, please consider your "normal" working conditions prior to the COVID-19 outbreak."
1554	ab762	Deviations from the Original Study The original study recruited college students and ran them on site, whereas the replication study recruited people online and ran them online. Although we did not gather data on demographics beyond nationality, we assume a more diverse sample in most respects (e.g., age, education, region, etc).
g5m	9bme4	Deviations from the original study:Despite our attempts to replicate the data collection and measurement procedures reported in Bersani and Doherty (2013), sample sizes differed slightly for reproduction analysis subgroups. Specifically, for the 1997-2009 period, we retained n= 2,819 ever arrested individuals versus n= 2,838 in the original study. Alternatively, for the ever arrested and ever married sample, we retained n= 843 versus n = 813 in the original study. We were unable to determine the source of these discrepancies.The estimation method also differs from the original study in key ways attributable to statistical software differences(i.e., HLM vs. Stata). In particular, Bersani and Doherty (2013) employed a population-average model with robust standard errors,coupled withallowance for "variation between individuals in the probability of an arrest and age[but not age-squared]parameters,"which we interpreted as a random slope parameter. Unfortunately, Stata cannot readily estimate or retrieve population averages from subject-specific random coefficient models, although it is possible to produce crude point estimates without standard errors or test statistics (Szmaragdet al.,2013). For this reason, we report (i)a population-average model and(ii)a supplemental subject-specific random coefficient model where the slope on age is allowed to vary by subject.
944y	bc2v6	Deviations from the Original Study The Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC) is used as the data source rather than the SIPP [Survey of Income and Program Participation]. More years of data are included : the original study uses data from December 1995 to February 2000; the replication uses data from March 1995 to March 2010). There are some deviations stemming from the fact that the CPS includes different variables than the SIPP and constructs some variables in different ways. Union membership is only available for a small subset of respondents and hence cannot be used. Unlike the original study, employees of private non-profit firms are included in the analysis (in addition to the for-profit firms that were the sole focus of the original study) because the CPS ASEC does not allow to separate them. The CPS does not follow individuals across ASEC questionnaires (while the SIPP is a panel). Clustered standard errors are used rather than standard errors allowing for autocorrelation given no panel data are available

g79m	vspzr	Deviations from the Original Study 1. While the original study used data from Israeli respondents, the replication data is from Indonesian respondents. 2. The original study uses eight adjectives for each personality factor (40 questions in total), while the IFLS5 uses three adjectives to assess each personality factor (15 questions in total). Moreover, the Likert scale changes across studies: in the original study the scale ranges from 1 to 9, but in the IFLS5 the scale ranges from 1 to 5. 3. In the original study, the authors consider the most prevalent diseases in Israel: diabetes, cancer, cardiovascular disease, respiratory disease (asthma, bronchitis, and chronic obstructive pulmonary disease), neurological disease (Parkinson's, multiple sclerosis, etc.), musculoskeletal complaints (neck pain and pain in the shoulder region or lower back), and rheumatic disease (e.g. rheumatoid arthritis). However, in the replication, it is assumed that the health conditions listed in the IFLS5 "Chronic Disease" modules represent prevalent health conditions in Indonesia. Using these diseases seems an appropriate approximation of this control variable in this context. 4. The original study contains measurements at two time points, though the focal analysis being replicated is restricted to the cross-sectional relationship at baseline. Additionally, the IFLS is a panel study, but only wave 5 is used in this replication, for two reasons: (1) limiting the replication analysis to one observation per individual mirrors the original 6 analysis more closely; and (2) measures for the focal independent variable (neuroticism) were not introduced until wave 5 of the IFLS.
k6y7	nuzh2	Deviations from the Original Study For one survey of the original dataset (Kenya 1988-1989), the key variables were unavailable. This is a small part of the dataset – 41 surveys were included in the original paper and only one of those surveys has not been included in the replication dataset.
95my	uc5zb	Deviations from the original study: 1. The original study uses the Comparative Manifesto Project (CMP), from 1945 to 1999, and includes the following countries: Australia, Austria, Belgium, Canada, Denmark, France, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, Netherlands, Norway, New Zealand, Portugal, Spain, Sweden, Switzerland, and the United Kingdom. The CMP is also used in the replication; however, more recent years (2000-2019) are added to the original sample, and nine more countries: Bulgaria, Croatia, Czech Republic, Estonia, Finland, Hungary, Latvia, Poland, and Slovakia. 2. The original study does not indicate the source for the type of electoral system. In the replication, the source is the Comparative Political Data Set (CPDS). 3. There is a difference of nine extra elections between the reproduction analysis (N=284) and the original study (N = 275). The difference can be due to an update of the CPM dataset; or because the CPDS dataset includes more elections than the source used by the authors.
y496	9t4nh	Deviations from the Original Study 1. Although both studies conducted a randomized control trial in Kenya, the study districts and the study period are different. The data for the replication was collected six years after the data was collected in the original study. 4 2. A deviation of the replication data from the original study is that the intervention is a combination of a free diagnostic test and an ACT voucher conditional on testing positive. The focal analysis is the effect of any ACT subsidy on ACT take-up, regardless of whether the illness is malaria, and regardless of whether it was tested. The replication data cannot separate the effect of the ACT subsidy from the free and positive rapid diagnostic test (RDT). 3. Following Cohen et al (2015, pages 622-624), the control variables are selected based on characteristics that do not balance across the treatment and the control group. The idea is to avoid any confounding in the estimates due to a lack of balance across groups. Consequently, the control variables are different between the original and the replication study.
7g66	wmqfu	Deviations from the Original Study Participants in the original study were recruited through Amazon Mechanical Turk; participants in the replication study were recruited through Prolific. We added two items to the items in the original study, one hidden item probing attention (Socimpact_Green_6), and one item asking how much effort respondents had put into the questionnaire (Effort_1). Both items served the detection of potentially non-serious response patterns and are documented in the Data Analysis Script (SCORE Data Analysis Script - Bhattacharjee_JournPerSocPsy_2017_Br0x.docx updated). We restricted participation in the study to respondents whose first language was English, in order to ensure that "noise" to the language barriers was minimal.
329k	kcegt	The main deviation between the two datasets is that the replication study includes all Indian states, not just Bihar and Uttar Pradesh. The other potential deviation is the identification of Backward Agricultural Castes (BACs), which are important for testing the focal claim in the paper. There is no ready classification for this in ARIS-REDS, and will need to be coded by hand or manually against a reference that delineates caste identity into the groups specified in the paper. This is possible because ARIS-REDS contains detailed data on caste and sub-caste at the household level. Additionally, data on income and village caste/land-holdings were collected in different ways and thus not fully compatible, though they should approximate the same concepts.
yk16	ybzev	Deviations from the Original study As mentioned in the preregistration, there are a few deviations from the original study. First, our sample is composed not only by students, but by working adults. This is because the replication target sample was considerably high to attain with students at our own universities, so we had to resort to an online panel to recruit the necessary number of participants. Second, the measurement of field of study (which is not necessarily important for our replication claim, but was included in the study for the sake of completeness) had to be modified to encompass not only the field of study, but also the field of work. Participants were asked to indicate which topics (from a comprehensive list of 28 topics) are important for their current studies/work. In addition, we note that the Political Knowledge scale had to be adapted from a German political context to an American political context to fit our sample of participants.

247z3	5r28d	Before summarization of claims, we will briefly describe the original study and our DAR attempt. Vadillo et al. (2016) attempted to analyze studies covered by Vohs (2015) in her review. Vadillo et al. (2016) argued that “vote-counting”, employed by Vohs (2015) in her review defending money priming, is not an appropriate technique to argue for the existence of the effect and that money priming experiments (in general, but especially studies covered by Vohs, 2015) provide evidence that is questionable. In particular, Vadillo et al. (2016) suggested that “evidence invoked by Vohs (2015) to support the robustness of money priming is compromised by selective reporting and other questionable research practices” (p. 656). For this purpose, Vadillo et al. (2016) decided to analyze: A) two classical money priming studies - Vohs et al. (2006) and Caruso et al. (2006). B) studies covered by a review of Vohs (2015) listed in Table 1 - Effects of Money Priming on Performance Measures Following Vohs et al. (2006) and C) studies covered by a review of Vohs (2015) listed in Table 2 - Effects of Money Priming on Interpersonal Measures Following Vohs et al. (2006). Claims selected by SCORE were related to points A to C and to specific meta-analytic techniques - Egger’s Regression, Funnel plot, Test of excess significance, and P-curve analysis. Thus, assuming that it should be possible to conduct DAR and that replication study of primary data is not the best solution with regard to the specificity of selected claims, we decided to find a different meta-analytic dataset that covered studies dedicated to money priming literature. For this purpose, the meta-analysis provided by Lodder et al. (2019) was considered the best candidate for DAR as this dataset belongs to the currently most recent and comprehensive meta-analysis on a money priming effect. Moreover, Lodder et al. (2019) provided open data that could be analyzed. These data allow for the analysis of Vadillo et al. (2016)´s focal points but with a different dataset. Although Lodder et al. (2019) focused on the same topic and covered similar studies (in fact, the is an overlap of studies – see “Disclosure table” on OSF for further analysis), data are not the same as in Vadillo et al. (2016). In particular: A) The analyzed data are not the same as effect sizes were computed by different authors. B) Besides computation of effects, inclusion criteria could differ for the original and DAR replication dataset. For example, when more than one DV was found in an article, Lodder et al. (2019) selected higher effects, while Vadillo et al. (2016) used first results (and used robustness analysis in some types of analysis). C) Vadillo et al. (2016) did not find all the intended data. In this case, the authors were contacted. However, even after attempting to contact the authors, they stated that “we were unable to access some of the unpublished studies included in Vohs’s Tables 1 and 2. Consequently, these could not be included in the analysis” (p. 659). Similarly, Lodder et al. (2019) were also not successful to include all existing (published and unpublished) effect sizes but made their best effort to do so (even mentioned Vohs in acknowledgments for her assistance in contacting the authors to gather the unpublished studies). Thus, the dataset of Lodder et al. (2019) allows corroborating the claims related to Vadillo et al. (2016) on a different dataset. Furthermore, besides a more direct attempt (specific claims as selected by score related to various sources - Vohs et al. (2006) and Caruso et al. (2006); studies in Vohs (2015) Table 1 and 2), it also allows us to conduct more conceptual sensitivity analysis where all original money priming studies are included (e.g., not only included in two classical studies dedicated to the money priming effect, namely Vohs et al. (2006) and Caruso et al. (2006); and studies covered by Vohs (2015) in Table 1 and 2; but all studies that are relevant to the topic of money priming and were covered by Lodder et al. (2019). In the next section, we will cover claims according to the analysis they refer to. Furthermore, we will also compute various sensitivity analyses to ensure that our results are not limited regard to errors in syntax or package idiosyncrasies. Additionally, in a general discussion, we will include a general sensitivity analysis to examine the generalization of the main idea of claims and all available studies in Lodder et al., 2019 (except replications and pre-registered studies as these studies should not be biased by selective reporting and problematic research practices).
288ok	syq4e	Deviations from the original study: 1. The original study includes 18 Organization for Economic Cooperation and Development (OECD) countries from 1970 to 2003. The data used in the replication contains 16 OECD countries from 2004 to 2018 (including 10 countries from the original study). Countries in the original study: Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Ireland, Italy, Japan, the Netherlands, New Zealand, Norway, Sweden, Switzerland, the United Kingdom, and the United States. Countries in the replication: Australia, Austria, Czech Republic, Denmark, Finland, Germany, Hungary, Italy, Japan, Korea, New Zealand, Norway, Poland, Portugal, Slovak Republic, and the United Kingdom. All the new countries belong to the “North” category (https://en.wikipedia.org/wiki/Global_North_and_Global_South). 2. The data sources are the same as the original study, except for the trade flows, that are obtained from UN Comtrade, instead of the OECD International Trade by Commodities Database. However, because both dataset measures trade flows across countries, the amount should be similar and therefore the impact on the replication results should be minimal.
65o96	e9j6y	Deviations from the original study: 1. The original study does not list the National Longitudinal Survey of Youth 1997 (NLSY97) rounds they use. Being published in August 2012, round 15 (corresponding to the year 2011) is presumably the latest wave that would have been available at the time of the original study. The data used in the replication is from the same source as the original study but using rounds 16 to 19 (corresponding to the years 2013, 2015, 2017, and 2019). However, given the panel nature of the data, it is also necessary to use rounds 1 to 15 to recover information such as respondent's fertility expectations by age 20; or to reconstruct the complete college attendance history for each person. 2. It is not clear from the study if they use the cumulative or panel sampling weights, or whether they use separate weights for each survey year or just weights from a single year. For the analysis, cumulative weights in 1997 are used. 3. The original study is not clear if they exclude respondents based on pregnancy or childbirth prior to leaving high school. On page 869 they explain that: “[...] we used a sample of 7,838 NLSY respondents who did not marry or have a child prior to leaving high school [...]” but on the same page, they also state that “We excluded those who were married or became pregnant prior to leaving high school [...]” (page 869). For the replication analysis the exclusion is based on the pregnancy variable because it should be a broader category that includes the childbirth event. 4. For Trace number 5 - Raley_JournMarFam_2012_D2LY_65o96, due to issues arising when combining a multiple imputation technique in a discrete-time proportional hazard model using logistic regression in Stata, a Cox regression is implemented.
308	ftv7q	Deviations from the Original Study: Unlike the original study, we collected data in an online setting using the software Inquisit Web (Version 6) by Millisecond rather than an offline setting in the laboratory. In addition, the replication deviated with regard to the sample and expanded the sample from Canadian psychology students to German speaking male and female adults between 18 and 30 years of age.
3053	b79mc	With regard to data collection, the only deviation, to the best of our knowledge, there is only difference between the original and replication studies. Specifically two constructs,(Cooperative Orientation and Competitive Orientation) were officially deleted from the PISA student questionnaire after 2003 and, thus, did not appear in the PISA 2012 data collection effort.Therefore, a replication attempt would need to be followed through without these two constructs. This could have implications on the replication attempt, as these independent variables were significant, however, their standardized effect sizes are considered very small (< 0.1). Furthermore, given the fact that these constructs were deleted due to potential invalidity, the results from the original study concerning these independent variables might also be not totally understood/valid.Consequently, it is expected that the non-availability of these independent variables for the replication attempt is minor due to potential invalid previous results.
286	sgnw9	While we did not deviate from our preregistration, we used Prolific to distribute our study via Qualtrics as opposed to Amazon’s mTurk platform. Additionally, the original study had 150 participants whereas this replication had 722 participants (out of the full 844). Our original plan was to have a 5% maximum exclusion rate but due to our strict exclusion criteria, we ended up with an exclusion rate of 14% resulting in our 722 participants.
mk67	579wh	Deviations from the Original Study: We had one key measurement difference in the replication. Namely, our measure of ‘Percentage of polling centers closed’ differs from Weidmann and Callen’s (2013) analogous measure of ‘Number of closed polling stations.’ We deviate from their measure for two reasons. First, data on the number of planned polling stations is unreliably measured; we have more confidence in data on planned polling centers. Second, normalizing by the number of planned centers adjusts for underlying district population differences as reflected by number of polling centers. Weidmann and Callen (2013) did not normalize polling station counts because each station is limited to collecting 600 ballots, so normalization was inherent in the measure. This is not the case with polling centers, which can support variable numbers of polling stations. We therefore normalize the measure as a percentage to make the measure meaningful in context. Analytically, while the main replication Models 1 and 2 were estimated similarly to the original, we estimated Models 3 and 4 to account for multiway clustering to account for both regional command (same as original) and election cycle (unique to this replication due to combining data from the 2014 initial and runoff elections).

65gm6	b2xma	The analysis uses entirely new data. The sustained growth measure deviates slightly from the original in that it will uses recent observations even if there are less than 5 subsequent years (e.g., post 2016). In the additional analyses section, I do a robustness where I restrict to data from the 2016 and earlier period to address this issue and the results are similar. The set of countries may also differ as the Afrobarometer survey expanded after Rounds 1-4. Again, in the additional analyses section, I do a robustness where I restrict to data coming from countries available in Rounds 1-4, and the results still fail to replicate.
41k2	gzwhe	Deviations from the Original Study 1. The original study relied on two student surveys that were collected at two different time points, 1994 and 1998. The data used in the replication are from the same sources as the original study but for survey years 2003 through 2006. 2. In the original study, “Parental education” is used as a proxy for socioeconomic status (SES). In the replication study, the Parents' total annual income is used as a proxy for SES. 3. In the original study, “Cross-racial interaction” (CRI) is a composite of student responses on five items. However, as it is explained in the Preregistration, in the “Data collection procedures” Section, the sub-questions that make up the CRI composite were not all fielded on all survey years after 1998.
2wyw2	59vgy	Deviations from the original study: 1. The original study relied on data from the fourth wave of the Indonesian Family Life Survey (IFLS), while the replication study uses the fifth wave. 2. The distance to the health center was used as an explanatory variable by Kim & Radoias (2016) in the main specification. However, in the replication sample, the variable contains serious limitations. For example, it was only collected from respondents who had visited a medical provider in the last four weeks. Further, only respondents who knew the distance have a value for this variable. Due to the issues with the distance variable, two probit regressions are estimated (i.e., one with and without the distance variable). This replication study will prioritize a probit regression without the distance variable as the final test for the focal hypothesis. The reason is to reduce the limitations arising from the measurement of the distance variable. However, a probit regression including the distance variable is also obtained.
m63	bnauj	Deviations from the original study: The biggest deviation that this study makes from the original is a shift to online data collection. This was done in response to the COVID-19 crisis, which shut down our in-person data collection. Fortunately, there is previous work showing that the SNARC effect is replicable with online participants (Cipora, Soltanlou, Reips, & Nuerk, 2019). This study found that the effect size of the SNARC effect was comparable to lab-based studies. In fact, the target of this replication (van Dijck et al., 2009) was one of the studies that served as a comparison in the analysis and the effect size in the online study was larger. Our own online sample found a similar result, with larger SNARC effects across all conditions. This finding should not be too surprising, given the existing body of work showing that the quality of data collected online is comparable to the quality of data collected in the lab (e.g., Buhrmester, Kwang, & Gosling, 2011; Germine et al., 2012; Crump, McDonnell, & Gureckis, 2013; Hauser & Schwarz, 2016; Hilbig, 2016). Of particular relevance is a recent methods paper reporting successful online replications for numerical cognition research specifically (Kochari, 2019). That paper used the same software used to program this study. Online studies have sufficiently accurate response time measurement capabilities for studies with randomized assignment like this one. (There is some concern about non-randomized designs because the specific device that a participant uses to complete the study may affect the measurement of response time, and device ownership may be correlated with demographic factors.) There is a large body of work showing that response time measurements for online studies, including specifically the software used to program this study, has similar variance to standard laboratory-based software (e.g., Anwyl-Irvine, Dalmaijer, Hodges, & Evershed, 2020; Bridges, Pitiot, MacAskill, & Pierce, 2020; de Leeuw & Motz, 2016; Reimers & Stewart, 2015).
7g95	4bckg	Deviations from original study: Tomz and van Houweling (2009) had each respondent complete mul- tiple candidate selections, including iterations that allowed respondents to familiarize themselves with the attributions of policy positions to candidates. This replication, however, asked each respondent to only make one candidate selection.
165m6	4gd23	Whilst a relationship was found between horizontal collectivist orientation and vividness recall of socialized power experiences, this was not related to ethnic/national belonging, contrary to what was inferenced (but not statistically confirmed) by original authors. Thus, in the generalizability study, one third Asians (India/China), one third Westerners (US/UK citizens) and one third Greeks were recruited to guarantee cultural orientation diversity in the sample and further explore the supposed ethnic nature of cultural orientation.
786	jqrde	Deviations from the Original Study The variable “credit standards” is taken from the same source the original paper used (“the Federal Reserve Board’s survey, ‘Senior Loan Officer Opinion Survey on Bank Lending Practices’”) but The variable “high-yield (HY) CDS-bond basis” is constructed from the two variables – (1) the yield spread of corporate bonds (available from Bloomberg) and (2) the CDS (Credit Default Swap) spread (available from Federal Reserve Bank of St. Louis), which is a different source than used by the original paper The original sample covers 2005 to 2009, while the replication dataset covers August 2012 to July 2020.
2k5g2	ck8aq	In some instances, the original article did not specify the variable used with sufficient detail to remove all uncertainty regarding the variable to choose in the Americas Barometer dataset (e.g., “iron hand” or “socioeconomics status”). Otherwise, there was no intentional deviation from the original article.
5zg9	6gpkx	Deviations from the Original Study: Unlike the original study, we randomized item order within the presentation of the behavioral inclinations and the DGS and MVS scales.
234w6	tv4se	Deviations from the Original Study The original study does not report which language was used in the survey and we do not have access to the original study materials. The measures reported in the study are available in English and we used the English version (Irish people are typically fluent in English – first language). In the original study, data was collected using a paper and pencil questionnaire and we collected the data using an online questionnaire. The recruitment in the original study was done exclusively via the preschools' administration. We sent out the initial recruitment email to preschool administrators and asked them to share the details with their staff but we also recruited participants via snowball sampling. Data collection took place in Spring 2022 during the COVID-19 pandemic. At this time in Ireland most COVID-19 restrictions were lifted (i.e., all retail, bars, restaurants, nightclubs are reopened, no mask mandate). Indeed, early childhood education centres (i.e., preschools) had been open since January 2021, and all other education settings have been fully open since April 2021. As such it is not expected that the COVID-19 pandemic will have had any meaningful impacts on baseline affective states but it is worth noting that this remains a possibility.
9k2y	fpku5	Deviations from the Original Study There were no major deviations in the variables taken from the fifth wave (2016) compared to previous waves.
y2312	hefyp	Our generalizability study differs from the original paper in the following ways: First, we conduct it online instead of in person, but the essence of the games stays the same. In this regard, our study directly speaks to the original claim 4 as it measures the same thing. Second, we focus on the general population in the United States while the original study ran the experiment with students from three different universities in India. The only criteria for inclusion in our study are that the participants need to be in the United States and between the ages of 18 and 65. Thus, this study speaks to the generalizability of the original claims and their external validity.
23w12	xyrv6	The original study was conducted with Israeli active-duty male military members. This replication deviates from the original in that we will use U.S. males with military experience (active duty or veteran).In our preregistration, we acknowledged that certain differences in the original study’s sample and our proposed sample may result in different observations of effects. For example, we noted how it would be possible the effect size between openness to experience and leadership experiences will be larger in the American population due to selection effects related to the voluntary nature of American military service (Israeli military service is compulsory). Thus, higher baseline levels of early life leadership experiences maybe present for the American sample compared to the Israeli sample, possibly resulting in a stronger observed relationship (i.e., greater than $\beta = .34$). This was not the case, however, as the effect size we found (unstandardized $b = 0.164$; standardized $\beta = .247$; $p = .001$) was similar to the coefficient found in the original study(both effects were medium in magnitude and positive in direction). We have no reason based on our results to believe the difference in sample populations meaningfully influenced our findings based on our results.

318z	xe3d2	Deviations from the Original Study The original study was conducted in person and the replication study is online. The method of compensation differed between the original and replication study. In the original study, participants were compensated with 20USD for a larger set of experiments including this one. In the present study, participants were compensated with 10CAD Amazon gift cards for this experiment only.
1yz2	fr6g8	Deviations from original study. The main differences between the original study and replication study consist in the sample's characteristics and setting of data collection. While in Usta & Häubl (Study 3) (2011) the sample consisted in thirty-six residents of Canada (21 women, 15 men; median age = 40.5 years) recruited through a university-run online participant panel, the replication sample presents a more equal distribution of participants relative to gender and age. Moreover, the replication sample is formed by British and United States citizens, whereas the original study recruited Canadian residents. More details regarding the characteristics of the replication sample can be found in the Replication Result section. Finally, although in both studies participants carried out tasks online, the replication study was conducted on a crowdsourcing platform which recruits participants from the general population (and not just university students) who could also carry out the study on their phones and tablets (and not just personal computers). The crowdsourcing platform chosen for data collection (Prolific.co) demonstrated to be a highly effective tool for the goals of the replication: recruitment and collection of data was very speedy and participants demonstrated high engagement to the tasks of the study. Lastly, a final deviation from the original study was due to the power analysis calculations, which rendered a significantly larger sample size for the replication study (with a final total of 274 participants after second round of data collection, compared to the 36 participants of the original study).
24812	jtd7n	The replication data comes from Youth Risk Behavior Surveillance System (YRBSS) survey. The 2015 wave was chosen for this replication effect as it was closest wave to the year of data collection in the original study (2008-2009) that included questions on sexual identity(i.e., previous waves of YRBSS did not ask about sexual identity). The respondents in the replication dataset were students from 125 schools in the United States. The YRBSS was developed in 1990 to monitor health behaviors that contribute markedly to the leading causes of death, disability, and social problems among youth and adults in the United States. YBRSS has been conducted biennially since 1991 and includes representative samples of students in grades 9–12. The sample is therefore slightly older than that in the original study where students in grades 7-12 were invited to participate. Another key difference between the original study and replication study is that the original study recruited opportunistic samples within the state of Wisconsin only and not a nationally representative US sample. Age and geographical differences may well affect the percentage of youths reporting to consider suicide, but these should still provide a good faith test of differences between sexual minority and straight youth. The replication dataset contains 15,624 respondents. After excluding those with missing data on any of the two key variables (sexual identity and suicide ideation), the final dataset contains 14,548 respondents (8.4% sexual minority youth, 91.6% straight youth; note in the original study this proportion was 4.9% sexual minority youth, 95.1% straight youth, therefore proportions are broadly similar with the replication dataset having a larger sample of sexual minority youth).
2g79y	6rpxu	The original study uses the National Social Life, Health, and Aging Project (NSHAP). The data used in the replication is from a different source that contains similar variables: the Midlife in the United States (MIDUS Refresher). More specifically, the MIDUS Refresher 1, 2011-2014 (ICPSR 36532). In the original study, the unit of analysis is the person-month, while in the replication the unit of analysis is the individual respondent. As such, the event history analyses described in the original study won't apply to this replication. Despite this difference, as in the original study, logistic regressions are employed; and it is possible to evaluate how marital duration and age impact sexual inactivity among older married people, which corresponds to the research claims of interest for the replication.

Table S5. Deviations from preregistered replication designs reported by replication teams

Project ID	OSF ID	Reported deviations from preregistered design
99kg	tb2up	Deviations from preregistration: None.
8z7g	dp4xv	Deviations from the Preregistration In the preregistration, we did not specify that we would test for distinguishability of dyads, as Campbell et al. did in the original study. However, when we conducted the analyses, we tested for distinguishability by imposing constraints on the a and b mediational paths and comparing model fit for this constrained model to an unconstrained model where men and women's a and b paths were freely estimated. Just as Campbell et al. found, imposing constraints did not significantly change model fit (indicating empirical indistinguishability between men and women), so we proceeded with the indistinguishable Actor Partner Interdependence model.
0y38	eq96h	Deviations from preregistration: For the current study, we did not include conference papers from the Academy of Management annual conferences 2016-2020, though we did conduct a search. The Academy of Management only publishes the first few pages of its submitted manuscripts. Thus, data were not available to extract from these papers.
y60	n8tds	Deviation from Preregistration The sole deviation from the preregistration was a minor deviation in analysis code to conform to the SCORE requirements. As such, the t-test coded as the focal analysis was changed from one-tailed to two-tailed, and an additional line of code was added to provide the Cohen's d effect size as part of the output. All deviations are therefore captured below and in the revised analysis code document (focal_revised): t.test(Converted_Data$fWHRPercent, mu = 50, alternative = "greater") to t.test(Converted_Data$fWHRPercent, mu = 50, alternative = "two.sided") cohensD(Converted_Data$fWHRPercent, mu = 50)
387	7edmy	Deviations from preregistration No changes or major decisions were made after the preregistration.
579	gu26j	Deviations From the Preregistration. The study materials deviated from the preregistered study materials in several ways. During cleaning of the pilot data, a variable name and a variable value were changed to match the contents of the data file, and the number of remaining participants at each stage was added in comments. During analysis of the pilot data, a frequency table was added to the analysis code. While testing the survey code, I discovered that the opportunity cost condition was not programmed. The Qualtrics code was updated to include the opportunity cost condition to match the design as specified in the preregistration. Before the replication data were cleaned, all changes made to the pilot data cleaning file were applied to the replication data cleaning file. In addition, the code was rearranged to construct all variables identifying excluded cases first and then apply exclusions (rather than construct each variable and exclude cases in sequence). While cleaning the replication data, I realized that the code used to identify cases with missing variables did not work as expected and modified it. At the end of Phase 2 data collection, before Phase 2 analysis had begun, I modified the focal hypothesis at the request of the SCORE team to clarify that in the opportunity cost condition, the correlation between connectedness to the future self and the decision to defer a purchase had to be positive (rather than merely larger than in the control condition). There were no other intentional deviations from the preregistration. This document contains links to the survey and code used in the replication, but the original survey and code remain available on the OSF project page.
393	mb5fp	Deviations from preregistration: COVID-19 pandemic related restrictions (study was conducted at the end of September) led to two minor deviations from the preregistration. (1) The data were not collected in 2 computer labs (as preregistered), but just in one of them. As preregistered, the participants were seated in separate cubicles, using standard computer equipment setup. (2) We originally planned to do code debugging and consequent code modifications using datafiles with reshuffled condition codes. However, due to data collection constraints, the analyst had access to the participant's datafiles, and blinding was not feasible. Ultimately, this had no bearing on the results because no substantial modifications of the preregistered script were necessary (also, the sampling method did not provide reasons for participant exclusions). No substantial changes to the preregistered analysis script were carried out. Only the following was changed. (1) In two instances, wrong column numbers / variable names were used. (2) A minor edit to the part of the code used to check the number of included trials after exclusions was done. This had, however, just a diagnostic function for the data wrangling. (3) To estimate the effect's R2 in the mixed-effects model, the default standardized generalized variance method did not converge, so we used the Nakagawa and Schielzeth method. (4) For the sensitivity analysis that excluded participants failing the manipulation check, we corrected a coding error, due to which the same response option was considered correct in both conditions (it was, in fact, different). None of the changes to the code had an effect on the results of the focal hypothesis test. The edits are documented in the code (https://osf.io/mn3y5/), denoted in comments as „CODE EDIT". There were no other deviations from the preregistration.
1964	fgdxs	Deviations from preregistration: Session size: session size varied from 6 to 21. This was due to differences in sign-up and show-up numbers. This will have no effect on results, since the sessions were run online. Session size is an important feature in physical labs, since it increases anonymity (the larger the session, the lower the odds of interacting with any one person). Online participants did not know how many people were taking part in the same session as them. Due to a lapse in the parameter setting, participation payoffs were set to £3.00 in the software instead of the pre-registered £3.50. The show-up fee was never displayed during the decision stages, so this would not have affected decisions. We realised the lapse before payments were processed. We adjusted the payments by the required amount and informed participants.
746	jmdsy	There were no deviations from the preregistration including in recruitment, design and analysis.
z516	kbxnr	Deviations from the Preregistration The preregistration (https://osf.io/94r7k/) did not include analyses at different levels of mean life satisfaction to further clarify the interaction effects. We added these analyses which are analogous to those extracted from the original paper, using three levels of low (bottom 25%), medium (middle 50%), and high (top 25%) of mean life satisfaction.
69412	2zv6n	Deviations from the preregistration: There are no deviations from the preregistration.
2y182	cdrmk	Deviations from the preregistration: None.
8y1	w52zk	Deviations from the Preregistration Data for the stage two was collected in two batches; the first was inadvertently collected by leaving the survey open after completing stage-1 data collection, and the second was collected after the confirmation from the project coordinators for stage-2.
g7g1	y8rgt	Deviations from preregistration. During conduction of the replication, some diversions were necessary in order to complete the study. In particular: In the sociodemographic section of the survey, three items were modified from 1st year undergraduate, 2nd year undergraduate, 3rd year undergraduate to 1st of post- secondary education, 2nd of post-secondary education, 3rd of post-secondary education. This change was introduced to render more universal items relative to educational level (considering the international nature of the cohort). It was determined to use κ^2 (or weighted kappa) instead of κ as this calculation of interrater reliability is more suitable for the rater scheme used in the study. Cohen's κ takes into account disagreement between the two raters, but not the degree of disagreement. This is especially relevant when the ratings are ordered, as in the case of the strategy used in the original study and in the replication. The weighted kappa is calculated using a predefined table of weights which measure the degree of disagreement between the two raters, where the higher the disagreement the higher the weight. Weighted kappa was calculated through medcalc. Basic κ was also calculated. These outcomes can be found on the OSF project site. Considering the problematic nature of the dataset, due to the dubious quality and authenticity of the data collected from the Indian subgroup and the low number of individuals (N = 14) assigned to the VI cultural orientation, H* was analyzed and confirmed via simple linear regression and hierarchical multiple regression.

21z56	r58ch	A deviation must be noted in the calculation of WOA. During the calculations, it turns out that it was possible to have a divide by zero error in this calculation as the advice and the initial guess could be equal. This would create a divide by zero error when calculating the average of all of the weight of advices for all of the questions. To solve this program, any divide by zero errors were replaced by 0 since 0 is interpretable as advice had no influence on final estimate. Also, even though the paper discusses the WOA measure as if it is always bounded by 0 and 1, that is not correct mathematically and, in the dataset, there are some trials where the WOA falls outside of 0 and 1. Those are included unchanged in the dataset as the paper did not mention any kind of truncation nor was any kind of truncation preregistered. Otherwise, there were no other deviations.
21487	yd9bg	Deviations from the preregistration: No deviations from the preregistration were necessary.
5g9	cdqks	Deviations from Preregistration - The originally registered sampling plan read “We aim to collect a representative sample of Dutch citizens, based on gender and age quotas.” Instead, we have collected a representative sample of UK citizens, based on gender, education and age quotas. This is different from the original study, as they recruited undergraduates. We argue that a representative sample is either an equal or even stronger sample than a student sample for this study. Furthermore, given that the study has been conducted in the US, a UK sample is arguable closer to the original study than a NL sample. - The originally registered sampling plan read “We will be using a 1 to 1 Euro Dollar exchange rate. We used Euro instead of Dollar given that the replication will be held in the Netherlands and following the exact exchange rate would make it unnecessarily difficult.” Given that we now have sampled among UK citizens, we used a 1 to 1 Pound Dollar exchange rate. Given that the range goes from 0 Pounds to 3 Pounds we would argue this is appropriate and following the exact exchange rate would make it unnecessarily difficult. - In the preregistration we noted “Participants will also be removed if they have not answered all questions. All questions will be mandatory so this means that we will only remove participants if they stop the game prematurely.” and “Participants with missing data will be ignored from the analysis where these data are needed. We will test whether or not using these data affects our hypothesis tests.” There were a very low number of missing variables for people who ‘completed’ the questionnaire. However, the missings were not needed for testing the focal hypothesis. We therefore used all 545 (first wave + second wave) who completed the questionnaire for testing the focal hypothesis, irrespective of whether they had some missing data on other variables. We also tested whether removing participants who failed to answer all questions affected the results of the focal hypothesis test. From the (n=545), a total of 3 participants had missings. Reanalysis without the participants who had missings, the results for the focal hypothesis remain statistically insignificant F(1, 6500) = 1.45, p=0.229, n=542.
68	g4htw	Deviations from the preregistration: None.
9kgg	n4kby	Deviations from preregistration - In the preregistration, we estimated a proportion of 9% inattentive responders and thus planned to collect a sample of N = 525 to achieve an analytic sample size of N = 477. The proportion of inattentive responders turned out to be slightly smaller than expected, thus we were able to reach the analytic sample size of N = 477 by collecting just 505 participants. - The code of the analyses was adapted to an update of the SemNeT R package. For example, since version 1.3.0 function partboot was replaced by a function called bootSemNeT, with a slightly different structure. We used the updated package and functions.
657	snrcv	Deviations from Preregistration: The current effort deviated from the preregistration by revising the SCORE test (H*) to specify the expected pattern of results. This change was made after Stage 1 data collection prior to Stage 1 data analysis. Prior to the revision to the SCORE test (H*), the pre-registered H* was “The mean level of preference for part-time work will vary among the three clusters of job seekers (i.e., satisficers, free agents, and maximizers).” Further, we conducted additional cluster validation techniques (i.e., McNemar’s test, silhouette test) that were not specified in the preregistration.
05g8	y4tra	There are no substantive changes made from the pre-registration
46z8	tywer	While we did not deviate from our pre-registration, we used mTurk to distribute our study viaSurvey Monkey. Additionally, the original study had 106 participants whereas this replicationhad 262 participants. Similar to Netemeyer et al., we reverse-coded scores on the CurrentMoney Management Stress (CMMS) scale “so that both scales were coded with higher numbers being more positive” (Netemeyer et al., 2018, p. 82). In our pre-registration, we did indicate that reverse coding on the CMMS would occur.
887	s26p5	There were no major deviations from the preregistration. Our analysis script (analysis.Rmd) was updated (line 18) to read the actual Qualtrics data file (data_round1.csv) instead of the simulated data (data_simulated.csv). We also added a section (lines 123 to 128) in our analysis script (analysis.Rmd) to randomly select five participants to receive the $20 bonus. Finally, we also added a section (lines 130 to 135) to save the cleaned/preprocessed dataset (data_cleaned.csv) in the Data folder.
6o946	4rch2	Deviations from the preregistration: There are no deviations from the preregistration.
yzm6	7xwyg	none
89z7	te7z6	Our pre-registered target sample was 29 participants, but we accidentally collected data from 30 (after exclusions). This happened because a slot was automatically reopened when a participant unexpectedly timed out on the experiment on Prolific. We had already opened a replacement slot for that participant. We included all 30 subjects who met the inclusion criteria in the analysis.
k17	uwqs5	Deviations from the Preregistration: The study deviated from preregistration in the recruitment method. We initially planned to contact kindergarten administrators to ask them to participate in the study. After agreeing, the administrators would have been asked to estimate the number of respondents in the organization and provide the emails of the employees who would potentially be interested in participating. We would then have contacted the employees with a recruitment letter. Additionally, this strategy would have been complemented with snowballing. Due to low response rates from the administrators, we changed the recruitment method to contacting kindergarten administrators and asking them to forward the email containing the link to the study (consent form) to the employees at their organization. The preregistered data processing script (k17_data_preparation_preregistration.R) contained one mistake that was corrected: in the original file, the columns containing PANAS-X items were renamed incorrectly (section “updating column names”), mismatching the items and the subscales. The renaming section was corrected so that the items match the subscales. The preregistered data processing script was changed in five instances: 1) the data loading part was added to refer to the raw data; 2) a new variable “Wave” was created to indicate the week of data collection (section “identifying participants present in all waves” in the k17_data_preparation_final.R script file; 3) a new variable “Age” was created by subtracting participants’ birth year from 2021; 4) the code was added to account for participant attrition across the time points and to remove participants who completed study more than once (sections “identifying participants present in all waves”, “selecting participants who completed the questionnaire”, “adding the number of the data collection wave to data files”, “Combining into one dataset”, and “Removing participants who completed the study more than once, keeping the first instance” in the k17_data_preparation_final.R script file). One change was made to the preregistered analysis script (k17_analysis_preregistration.R): the scaling of predictors and the outcome variable to obtain standardized (as opposed to unstandardized) regression coefficients. The replication was conducted during the COVID-19 pandemic, with data collection taking place in summer-fall 2021. The data collection period was interrupted by mandated closures of educational institutions, including kindergartens, and summer holidays. The data was collected from institutions that were open or open with restrictions at the time of data collection. However, the COVID-19 pandemic could have affected the baseline affective states, work- and home- related hassles, quality of weekend experiences, and the psychological separation between the work week and the weekend among the kindergarten teachers. The unique context of the COVID-19 pandemic needs to be taken into account when interpreting the results of this replication.

8g91	j24ts	Deviations from the preregistration: For reaction times in the lexical decision task, we have planned to delete raw reaction times below and above 3 standard deviations. Instead, we deleted mean reaction times below and above 3 standard deviations. We also deleted raw reaction times in the main task, which were below or above 3 standard deviations, which was not included in the preregistration. Mean reaction times were calculated without the outlying RTs.
y071	69zjt	Deviations from preregistration: 1. Instead of R(Studio), Version 3.6.1 we used Mplus 8.4 to analyze the data, given that R does currently not allow for multilevel ordered logistic regression with regular inference statistical methods. 2. We adjusted the chat room given that our initial approach violated Amazon Mechanical Turk's privacy notice. Previously, we planned that participants will insert their first and last name in the sham chat room. However, MTurk does allow participants to provide this private information (even it is not saved and other persons in the chat room were not real). We therefore changed the chat room so that we asked participants for a nickname or alias (e.g., "LuckyLuke41") instead of first and last name. 3. Data collection started later (i.e., from 5th of November, 2020) and took longer (i.e., 9 days until 14th of November, 2020) than proposed in the pre-registration. 4. The consent form was already presented when the participants entered MTurk so that participants were directly fully aware of all information regarding their participation (in the pre-registration we stated that the consent form will be only at the stage of the Qualtrics survey). 5. In order to achieve the analytical sample size of 573 or more (our final analytical sample size was N = 574) we had to sample data from 898 (not 688) MTurk Workers, given the higher-than-expected attrition rate (i.e., 36% rather than 20%). 6. The links to the Qualtrics survey (https://unigiessen.eu.qualtrics.com/jfe/preview/SV_7UQSBUYAvM6CWBD?Q_CH Formatted: English (United States) L=preview&Q_SurveyVersionID=current) and the sham chat room (https://score- Formatted: English (United States) chat.herokuapp.com/join/wtf78gs4ly/) needed to be updated. Formatted: English (United States) 7. We did not ask for MTurker's ID, given that it was not necessary (in contrast to what Formatted: English (United States) we expected before data collection) to gather this information. 7.8.We coded favourable temporal social comparisons as 0 (instead of 1, as noted in the pre-registration) and unfavorable temporal social comparisons as 1 (instead of 2). Similarly, we coded low competition as 0 (instead of 1) and high competition as 1 (instead of 2). We realized that the relationship pattern and the simple effects are easier to interpret that way (THIS IS AN ADDENDUM ON JULY 13, 2021)
1634	kw5tv	Deviations from preregistration none
23wwg	srk8q	Deviations from the preregistration: None.
19gz	xg6q5	Deviations from the Preregistration There are substantial deviations from the preregistration: The dataset (Final_Data_AnnualScreening.csv) contains firms which have less than 750 observations, so I excluded those as the original only included firms that have at least 750 observations. This reduces the number of firms included to 2142. The pre-registration called for running one regression for the whole sample. However, the original study ran separate regressions for each entity and then reported summary statistics (mean, median) for the t-values and p-values: Page 480: And page 481: " Accordingly, it was decided to deviate from the pre-registered regression and run separate regressions for each entity. The estimated focal coefficients from each of the 2142 regressions were collected into one sample, and then the mean of this sample was tested against a null hypothesis of 0.
m89	82z6a	Deviations from Preregistration The authors of the original study conducted the study in two phases where they recruited participants in person during class time for Phase I, provided them with prepaid envelopes, asked them to mail back their Phase I survey responses, and finally emailed them the Phase II survey upon return of the participants' Phase I survey responses. Given the pandemic of COVID-19, we modified data collection to collect data for both phases completely online. We contacted faculty from the College of Business to email our recruitment message to their upper-division 3 undergraduate and MBA students (the inclusion criteria). In addition, we made use of our university's listserv to send our recruitment message to all undergraduate and graduate students at the university. We were unable to specifically target upper-division undergraduate students and MBA students through the listserv method. Nevertheless, to accommodate for this issue, we created a pre-screener at the beginning of the Phase I survey to confirm that the participants met inclusion criteria before viewing the consent form. The pre-screener included an additional item to the original study, asking if the participant is an undergraduate upper-division student or MBA student and if so, which one specifically. The second item referred to the employment status (unemployed, part-time, full-time) of the participant, which was a demographic item from the original study. The original authors used Colquitt's (2001) seven-item measure for procedural justice. The following statement marks the beginning for each of the items, "The following items refer to the procedures used to arrive at your procedural justice. To what extent…" We were concerned that participants would not understand the meaning of 'procedural justice.' Previous attempts had been made to contact the original authors to clarify, among other things, how they presented the directions/items to the participants. We did not receive a reply, so per the suggestion of OSF, we modified the measure for better comprehension: • The following items refer to your organization's procedures and treatment of you. To what extent: o Have you been able to express your views and feelings to your organization? o Have you had influence over your organization's treatment of you? o Has your organization's treatment of you been applied consistently? o Has your organization's treatment of you been free of bias? 4 o Has your organization's treatment of you been based on accurate information? o Have you been able to appeal your organization's treatment of you? o Has your organization's treatment of you upheld ethical and moral standards? The original authors asked the participants to disclose their sex identity. We were unable to determine whether they accommodated for one's transgender identity and/or whether they meant sex assigned at birth. Therefore, we used best practices from the Williams Institute in our approach to ask the participants about gender identity. • How do you describe yourself? o Male o Female o Transgender o Do not identify as female, male, or transgender
4z32	fngwe	Deviations from preregistration Due to the COVID-19 pandemic all data were collected online using mTurk (see "Other" on preregistration), so we are unable to analyze any potential differences between online and lab participants. Second, instead of receiving $5 in mTurk credit, the budget was adjusted to pay participants between $2.50-$5.00, to accommodate fees and the need to discard subjects who did not pass the compliance and audio quality questions (see below). As described in the preregistration (see "Other"), we included four questions to assess subject's compliance with instructions and attention: 1) Subjects heard an audio file instructing them to answer "red" to the question "What color is the sky?" which allowed us to assess compliance and verify that they could hear stimuli, 2) Did you have technical difficulties? 3) How carefully did you complete this survey? and 4) Did you wear headphones? Subjects were asked to answer these questions honestly and told that they would be paid regardless of their answers. We excluded subjects who failed to answer "red" to first question (N=176), who reported having technical difficulties (N=3), or who admitted they were not careful (N=1). Because posted recruitment ads and instructions encouraged but did not require headphone wearing, we did not exclude participants on this basis, however when we conducted the analysis only including subjects who wore headphones (N=594) we observed the same pattern of results.
24716	tybxg	Deviations from the preregistration: None.
6m492	wygup	Deviations from the preregistration: No deviations from the preregistration.
21wg2	5xr92	Deviations from the preregistration: There are no deviations from the preregistration.
658m2	nbc6v	Deviations from the preregistration: There were no deviations from the pre-registration.
67z12	9hjxc	There were no deviations from the preregistration.
32mk	dyn6j	Deviation from preregistration There were no deviations from the preregistration in Stage One. Stage Two involved a 2nd pretest collection as the number of potentially eligible participants that emerged from the initial pretest were less than anticipated.

6m34m	2jmtf	Deviations from the preregistration: in the preregistration RD*DTP interaction term was included by mistake. It should be RD*FTI (as per claim 4 text). The results provided here are of RD*FTI interaction.
69y36	dtkjc	Deviations from preregistration: No deviations from the preregistration.
m707	qk34m	Deviations from the Preregistration There were four deviations from the preregistration. First, participants were invited to participate in wave 1 of the study only if they indicated in the Sona pre-screening that their racial background was White, which they did by checking a box from the following options: White, Black/African American; American Indian or Alaska Native; Asian; Hawaiian or Pacific Islander; Multiracial; Other. I pre-registered a plan to exclude participants from the final sample if they did not self-identify as White in wave 2. My intention was to allow for the exclusion of participants who may have accidentally selected “White” during the Sona prescreen. However, participants provided a variety of responses to the open-ended question about their race in both waves, including the following: American, Caucasian, Hispanic, Italian, Latinx, Middle Eastern, Puerto Rican, White, and various combinations of these. Given that all of these participants self-identified as White during the Sona pre-screen, I included all of these participants in the final sample. Data were excluded from one participant, however, who responded “Yes” to the wave 1 question “Do you identify as Black/African-American? Please answer yes if you are biracial or multiracial and one of those identities is Black or African-American.” In addition to conducting analyses on the full sample (N = 113), I conducted parallel analyses using a conservative sample that included only participants who specifically self-identified as White in either wave 1 or wave 2 (N = 92). All of the conclusions were consistent across both analyses. 7 Second, I pre-registered a plan to compensate participants for the second wave of the study with research credits or a $10 Amazon gift card. Ultimately, however, participants did not sign up to participate in wave 2 through Sona and instead participated in the study directly from the link provided in their email. Accordingly, all wave 2 participants were compensated with a $10 Amazon gift card. Third, I preregistered a plan to send two additional follow-up emails to participants after the first wave 2 email request. Toward the end of the data collection period, however, I sent two additional follow-up emails to participants who had not yet participated in wave 2. Fourth, I was asked by the SCORE team to report the results of the focal test (the effect of communication positivity on post-test anti-Black attitudes at low levels of IMS) even though my preregistered plan was to do so only if the Subjective ratings * IMS interaction was significant.
128g6	zjt47	Deviations from the preregistration: There were no deviations from the preregistered protocol.
6zoo2	gyse6	Deviations from the preregistration: There are no deviations from the preregistration.
6m396	atwck	Deviations from the preregistration: There were no deviations from the pre-registration. Additional analyses (optional) Manipulation check. We rain analyses to determine if the mood manipulation was successful. We found that for participants who thought that it was not moral to help build houses at the expense of helping a family m ember, there was not a significant difference between participants in the positive mood condition and the negative mood condition on valence, t(107) = 1.7217, p = .08801, nor on arousal, t(108) = 1.1745, p = .2428. This is in contrast to the original paper, which found a difference between participants in the positive and negative mood conditions on valence, but not arousal, for participants who decided to not push the man (trolley dilemma). For participants who decided it was moral to build houses at the expense of helping a family member, there was a significant difference between the positive and negative mood conditions on valence, t(195) = 3.656, p = .0003295, but not on arousal, t(198) = 1.6361, p = .1034. This is also in contradiction with the original paper, which found that among those who decided to push the man (trolley dilemma), there were no differences between participants in the positive and negative mood conditions on valence or arousal.
g77m	fqdh4	Deviations from preregistration: In our preregistration plan we planned to run a two-sided linear probability model without any control variables. Furthermore, we initially included public housing authorities run by county governments. To exactly replicate Table 1 in Einstein and Glick (2017), we exclude county government run public housing authorities, run a logit model, and include control variables. Tests of significance are two-tailed. We deviate from our preregistration in order to exactly replicate Table 1, but we subsequently follow our preregistration plan and estimate the ethnic differential using a two-sided linear probability model without additional covariates. In both cases we fail to replicate Einstein and Glick (2017)’s main findings that Hispanics are less likely to receive a friendly response when contacting public housing authorities.
369z	dm287	The preregistration did not mention removal of any outliers. However, a single extreme outlier (reporting a willingness to pay of £1x10^33) in the risky prospect group, who otherwise passed the comprehension checks, was found to be skewing the data. Exploratory analysis after removal of the outlier showed that replication of the claim was successful according to the SCORE criteria. COS team members who were blind to the results of the two analysis versions made the decision that the single extreme outlier could be removed, thus the decision was made to report the exploratory analysis as the replication claim test. When setting up the survey on the Prolific platform, we found that Prolific’s minimum compensation rate meant the minimum we could pay participants was £0.42 GPB, not the £0.25 GPB detailed in the pre-registration. This change was made.The pre-registration stated that the lottery condition would be coded as ‘1’ and the gift card condition would be coded as‘2’. In actuality, the gift card condition was coded as ‘1’and the lottery condition was coded as ‘2’.
65km6	wgznk	no deviations from the preregistration occurred.
935y	jx64p	Deviations from preregistration. In our OSF preregistration and the original article, Dumas & Perry-Smith (2018), it states the work absorption scale is measured on a 7-point scale with “1 = Strongly Agree, 7 = Strongly Disagree”. This is atypical Likert scale use and would actually result in a construct that is not work absorption but lack thereof. The correct scale anchors should be the opposite: “1 = Strongly Disagree, 7 = Strongly Agree” scaling. We implemented this change after corresponding with the original study first author who confirmed that their reporting was a typographical error. We attached a PDF copy of the communication with study authors for SCORE coordinators review. In the original preregistration a Phase 2 was planned should the effect not replicate in Phase 1. Upon collecting Phase 1 data, however, and look at the magnitude and direction of this effect size -0.01, which is not only opposite in magnitude to the original study but vastly different in size, we believe Phase 2 will not alter our replication result. SCORE coordinators concurred. No other deviations to report.
8m17	qu793	Deviations from the Preregistration Both participants that were run online took significantly longer than in-lab participants and were unable to finish all 144 trials due to connection issues. They both finished the majority of the trials and the analyses were run with and without their data. Although we secured a reliable internet connection, we realized that we could not rely on participants having the same, so we instead took extra measures to bring the remaining participants to the lab.
95y	mfpwk	Deviations from preregistration There was no deviation from the preregistration.
2g7z2	n6jca	Deviations from the preregistration: [After discussion with OSF personnel, we decided to not run the full ANOVA and run the focal test using a dummy coded regression instead of a 2x2 ANOVA. Additionally, for interpretation, we ran an independent groups t-test. This was partially due to the calculated power from the original replication attempt yielding a sample size of 516. In order to run this across both single and non-single women, it would have required a minimal sample size of 1032. Participants that entered that they were non-single or participants who did not enter the desired compensation values were excluded prior to data analysis (yielding our sample size of 528). We had to collect additional data. Despite the prolific filters, participants that were non-single (married/cohabitating) participated in the study. After an additional round of data collection, we ended up with a final sample size of 528 (269 in the anonymous condition and 259 in the disclosed condition). We changed from Qualtrics to Survey Monkey. Survey Monkey has the same randomization options and this would not have impacted the results of the study. Finally, we didn’t disclose how we were going to calculate the DV (desired income) in the preregistration. As the original author’s did, we calculated the midpoint of the range selected (if there was a range) and used the reported amounts of $62,500 for under $75,000 & 262,500 for over $250,000.]
675m9	u58za	Deviations from the preregistration: None.

2yg	2ujvn	Deviations from preregistration: Preregistration mentions statistical model as follows: “As per the description of the study and hypothesis: A country’s average score on the emancipative values index (‘RESEMAVAL’ - treated as random effect) will be positively associated with the protest level of respondents from that country. (…).” In fact, ‘RESEMAVAL’ is an individual-level variable and the the focal result should be for 'societal emancipative values'. In this replication it was coded as ‘emancipative_climate and is treated a s a fixed effect’. Thus the model description changes to: “As per the description of the study and hypothesis: A country’s average score on the emancipative values index (‘emancipative_climate’ - treated as fixed effect) will be positively associated with the protest level of respondents from that country. (…).”
67g8	8npm5	Deviations from preregistration: (1) The preregistered build of the experiment included an audio check. We modified this to be an audio/video check as a precautionary measure. (2) Minor changes to the preregistered SPSS syntax were made due to the fact that the ‘Finished’ variable provided by Qualtrics did not discriminate between participants who had genuinely finished and those who had been automatically exited from the experiment due to either (a) accessing it on a mobile device; or (b) failing the abovementioned audio/video check.
65wm6	qev3x	Deviations from the preregistration: None.
7976	ad9jm	The project was preregistered as having two stages of data collection, with additional data collected in Stage 2 if we did not replicate the key result in Stage 1. However, due to COVID delays in our completing of the project and time constraints on the SCORE project, we could not complete the Stage 2 data collection for this project. Thus, the only major deviation from our preregistration is that despite failing to replicate the key result in Stage 1 we did not collect Stage 2 data. Because this project involved significant interaction with, and touching of, human subjects, we faced a major delay in running participants and were only able to do so starting in late January of 2022. Thus, whereas our stopping rule was preregistered as continuing to stage 2 if stage 1 resulted in a non-replication of the key result, we were unable to collect Stage 2. An additional minor deviation is that while our preregistration and Data Format specification indicated we would record gender as M/F, we had several participants report they were non-binary, and so included Male/Female/Non-binary categories in our final data file. Gender is not relevant to any analyses in this project so this has no consequences for the test of the key hypothesis.
999g	n2m5w	Deviations from preregistration We used an additional independent sample t-test to analyse the data, after removing unrealistic extreme values (i.e., > $20). This resulted in a sample size of 147 participants. The additional test is identical to the planned test, both are independent sample t-tests with the same variables.
yk20	dpjef	Deviations from Preregistration The xtgls routine estimated in STATA has been augmented with the options “panels(hetero)” and “corr(psar1)” in order to control for serial autocorrelation in the residuals, spatial cor- relation, and groupwise heteroskedasticity. This resulted in an estimation routine closer to what the author outlined in the published article. Although more consistent with the text, this change did not alter the results of the replication, or the ultimate assessment of the replication claim.
4zy8	gaqbx	Deviations from the Preregistration In our dataset, seven participants were tied for 22.7-28.0 percentile, so there were 37 rather than 33 participants in the extrinsic group. The closest break point for the 75.0 percentile meant that there were 34 rather than 33 participants in the intrinsic group. This was the closest we could get to having 33 in each group without going under that required number.
2kkg2	vu8nb	Deviations from the preregistration: None
16yz	6aefs	Deviations from the Preregistration There are no known deviations from the original study.
zz26	udbnf	The two major deviations from the preregistration were discussed previously: 1. For class identification we stated we would conduct 1000 runs from the final class solutions, but due to logistics this number was reduced to 100 runs. 2. We did not actually specify thresholds for how to determine the presence of the target classes across the runs. Accordingly, we made the thresholds consistent with other specifications in the project (80% consistency). Although these deviations are limitations, the results provide clear evidence of the inability to reliably identify the target classes. SCORE Syed 8 There were four additional deviations related to the survey construction and administration: 1. The four eligibility criteria for the study—ages 18-30, speak English, live in the U.S., and race/ethnicity—were moved to the beginning of the survey so that participants could be properly screened. 2. In addition to the open-ended race/ethnicity question we added a checkbox question, which was necessary to determine participant eligibility. 3. At the recommendation of Qualtrics, an additional quality check item was added to the beginning of the survey. 4. The two attention check items were treated as indicators of “quality completes” by Qualtrics, so participants who failed were screened by Qualtrics not by the research team.
931g	9s4e6	Deviations from preregistration: We did not deviate from the pre-registration.
276	fq2vg	Deviations from preregistration. In Stage 2, we pre-registered to collect data from 512 participants. However, as the experiment was sent out to a larger participant pool, more people decided to take part in the time interval they were allowed to enter the experiment. Hence, Stage 2 analysis was conducted on 533 participants.
895g	fzkm3	Deviations from preregistration: In section 8 Study design of the preregistration file, it was mentioned in the slider task section that the subjects will see a number of slider bars positioned at 0 on their computer screen. In the replication study, the subjects saw one slider bar positioned at 0 on the computer screen at a time, rather than seeing 48 sliders altogether as indicated in the preregistration. All other aspects of the earning stage were executed as described in the preregistration.
20g	xkwj9	Deviation from Preregistration Although the stage 1 results did not show a statistically significant effect at $p < .05$ for the focal H* test, stage 2 data collection was not conducted due to ending of Phase 1 of the SCORE project and the unavailability of the additional funds needed to conduct stage 2. The preregistration specified a sample size of 144 to achieve an analytic sample of 122 based on the exclusion rate of the original study. Instead of recruiting all 144 participants and then excluding participants who did not meet the preregistered criteria, participants were informed of the exclusion criteria, which involved passing attention checks and meeting minimum performance criteria on the cognitive tasks administered, and told that if they did not meet these criteria they would be asked to “return” their survey or would have their submission “rejected” on the prolific.co platform so that a replacement participant could be recruited. Analyses of the attention checks and minimum performance criteria were then conducted in periodic batches while data collection was ongoing, and participant recruitment was continued until a minimum of 122 participants exceeded the exclusion thresholds in total. Therefore, although the stopping rule was not implemented exactly as preregistered, the same target analytic sample was recruited and none of the preregistered hypothesis tests were conducted until this analytic sample size was reached. As shown in the “prolific_export_exclusion.csv” file included with the raw data, a total of 210 participants began the study. Of these, 68 participants returned the survey (either because they chose not to complete the study or did not meet the exclusion criteria), 10 participants had their responses rejected because they did not respond to a request to return their survey after not meeting exclusion criteria, and four participants had the survey “time-out” for exceeding the maximum time allotted by prolific.co without completing the study. This left 128 participants who met the performance criteria for inclusion in the study. When assembling the data for analysis, one additional participant was eliminated for incomplete responses on the questionnaire measures, leaving a final analytic sample of 127. Thus, overall, although the specific preregistered exclusion criteria were not altered, they were procedurally implemented in a different way than described in the preregistration. The pre-registered data-cleaning syntax mistakenly did not filter out the practice trials on the cardsort task before calculating the number of trials that timed out before participants could respond on the cardsort task. This mistake was corrected before the focal analyses were conducted. The pre-registered analysis syntax mistakenly failed to eliminate incorrect responses when analyzing the logRT variable on the card-sorting task, which was part of the preregistered procedures. The correction was made to the final analysis syntax.The pre-registered analysis syntax produces raw slope coefficients. To better match the standardized ßs reported at places in the original article, this syntax was updated to z-score individual difference measures (i.e., Need for Cognition and BAS)and continuous DVs (i.e., logRT) around the grand mean for the focal analyses.

658g7	dkcu6	the name of the dataset in the model had a typo that was corrected.
1762	8rwkn	Deviations from preregistration There was no deviation from the preregistration.
6738	xn73j	Deviations from preregistration: There were two deviations from the pre-registration. First, the pre-registered inference criteria for a successful replication were updated (see “amended replication criteria” section above for more details). Second, the pre-registration stated that a 2 (decision frame: active, passive) x 2 (mood: positive, negative) Chi-Squared test would be performed as the focal test. However, a 2 (morality: moral, not moral) x 4 (mood/frame condition: positive/active, positive/passive, negative/active, negative/passive) Chi-Squared test was conducted.
546	f5esc	Deviations from the Preregistration There were no deviations from the preregistration.
2kgg2	3n2bp	Deviations from preregistration We planned to collect data from the University of Exeter FEELE lab first, and only then to move to Prolific.co. In actual fact, we only collected data from Prolific.co. We planned to use participants’ disciplinary backgrounds as an inclusion factor, such that only participants with disciplinary backgrounds in economics and business or the humanities completed the survey for their discipline. However, pilot testing for the power analysis revealed that the Prolific screening tool was not entirely accurate, allowing participants who had a background in business to complete the humanities survey and vice versa. To address this, we included additional screening questions at the beginning of each survey. We also asked participants to categorise their main disciplinary background (as either economics and business, humanities, STEMM or other) at the end of the survey. This provided the basis for a new exclusion factor. Specifically, participants who completed a survey in one discipline (e.g., business) but reported that their main disciplinary background was in the other discipline (i.e., humanities) were excluded from the analysis (but not the data set). We used the Prolific balanced sample option so that roughly half of the sample were male and the remainder female because of recent events that have strongly skewed the Prolific participant pool to female.
6zzo6	dnqrv	We calculated the proportion of excluded participants midway through the experiment and discovered that it was much higher than anticipated. The original study excluded 1 person for using external memory aids, but after collecting data from the first 222 people we found that 61 (27.4%) reported using something to help remember the objects. This threatened to exhaust our budget for the study before reaching our target sample size. We decided to modify the description of the experiment on Prolific to explicitly ask people to not use something to help them remember. In the second round of data collection, only 39 of the 352 people reported using an external memory aid (11%). Since condition assignment was cyclical, this change affected all conditions equally.
5796	hmune	Deviations from preregistration: The formal evaluation of both delta between the two G^2 values using a chi-squared goodness-of-fit test and phi to get the effect size were added after the preregistration.
z189	2f35a	Deviation to the Pre-registration Masking- Our originalPre-registration stated that masking would be done by qualtrics, however, this replication used Formr to administer and create the link for participants. Participants and the Experimenter were both still masked from the assigned condition.Deviation to the Pre-registration Stopping rule-Our original Pre-registration stated that once we had collected our stage 1 analytic sample size of 377 usable participants(recruiting 500 participants to account for attrition) we would stop and analyze the data.However, after collecting 504 participants and applying exclusion criteria, we did not meet the analytical sample size (377). We then ran 251 more participants to attempt to meet377. Running a total of 755 mturkers through the full experiment. However, you will notice our data has 899 respondents, this is due to the software, formr, maintaining participant data that was in fact a test run, or did not have the participant finish fully (which would lead to them being excluded from analyses). After exclusion criteria, we ended with 376usable participants. Our stopping rule also stated that if we did not meet the analytical sample size we would run “batches” of 50 until successfully reaching 377. Due to budget constraints, we were not able to run additional "batches" of 50 to attempt to recruit more participants after exclusion criteria.Deviation from the exclusion criteria-Our originalPre-registration stated that we had three exclusion criteria. However, since formr kept test runs or data run before the link was posted to mturk and after the last mturk session was completed, we excluded these cases as they were not legitimate participant responses.
24yw6	4usmq	Deviations from the preregistration: There are no deviations from the preregistration.
67y16	3u8x9	Deviations from the preregistration: There are no deviations from the preregistration.
3z5z	dw4mf	One deviation from the pre-registration was necessary. The pre-registration, as well as the original paper, indicated the use of panel corrected standard errors. There was no problem estimating results when using the required test sample for the preregistration. However, when the same code was applied to the full dataset, the models were no longer estimable due to computational singularity. This is often due to linearly dependent covariates. A correlation matrix, however, reveals that correlations between the independent variables were quite reasonable (highest was < 0.5). Removing the year-and state-specific intercepts allowed the model to be estimable, as did changing the panel corrected standard errors to Huber-White sandwich corrected standard errors. Turning back to the sample-based models used in the pre-registration, I re-estimated them first with only removing the year and state intercepts and then with only changing the standard error correction method. Removing the year and state intercepts had a much larger effect on the estimation of the coefficients, standard errors, and p-values than simply changing from panel-corrected standard errors to Huber-White. Thus, I decided to make that change. Thus, in a deviation from the pre-registration and the original paper, I use Huber-White standard errors instead of panel-corrected standard errors.
z4z9	jr239	Deviations from the Preregistration There were no deviations from the preregistration.
z591	d923e	Deviations from preregistration: We observed a substantially higher rate of exclusions than the original authors. In particular, where the original authors reported that 35% of participants failed at least of the comprehension checks, 58% of our participants failed. To adjust for this, after the first batch of participants, we recruited around twice as many participants in each additional batch to achieve the stage 1 and stage 2 samples of 110 and 246 participants. This resulted in a slight over-recruitment, of 113 participants for stage 1 and 247 for stage 2.
5z36	unj3f	Deviations from the Preregistration The final data analysis was conducted after the preregistration document was reviewed and accepted as final by the Action Editor, but before the project was registered on OSF. Therefore, although this project is technically classified as “registered following the analysis of existing data”, all of the data source identification, variable selection, data transformation, and analysis code was preregistered and approved by reviewers before the final data analysis was conducted and there were no deviations or modifications from the reviewed preregistration in how the replication analyses were performed. That is, all of the decisions made concerning the project were data-independent and were not altered in any way between performing the replication analyses and the final registration on OSF. Following the primary replication analyses, given the reversal of the interaction effect from the original study, several additional exploratory robustness analyses were conducted that were not initially preregistered. The first analysis involved additionally including the specific political party being rated as an additional random effect in the overall regression analyses. This addition did not change the results in any meaningful way. The full output from these analyses can be found at https://osf.io/q7yrv/. The second analysis involved conducting the entire replication study again with the CSES Module 5 data set. As explained in the preregistration document, this data set was also appropriate for testing the primary claim from Zakharova and Warwick (2014), but was less preferred than CSES Module 4 because the Module 5 data is still only in advanced release and has not been completely finalized. The R code for assembling, transforming, and analyzing the Module 5 data can be found at https://osf.io/g7wfs/, https://osf.io/6t42n/, and https://osf.io/hfyad/. Note that, due to conversion errors in the respondent ID numbers for Module 5, all instances of variable E1005, the respondent ID for the full study, that are found in the code was replaced with variable E1009, the respondent ID within the country-specific sample.

75g6	2wfgv	Deviations from preregistration: Only one change to the analysis script was made, namely removing all 'testdata' lines and only including 'fulldata'. Furthermore, p-value calculation was originally only added to the focal analysis, so they were added to the other 7 additional analyses as well. Some line numbers in the commenting were changed as well. This also revealed a typo in the code for additional analysis 6, hence: Error a6_totalvar <-a6_bvar_adjusted+a2_meanvariance_of_vars Was changed to: Error a6_totalvar <-a6_bvar_adjusted+a6_meanvariance_of_vars
6zz3k	5869t	Deviations from the preregistration: None.
67516	m8bpk	Deviations from the preregistration: Upon printing the materials to run the first session, I realized that I only uploaded the comprehension quiz for the TS treatment. The OS comprehension quiz used only Q1 from the uploaded questionnaire, and removed any mention of "second stage".
2g7ky	8mqzu	There are two deviations from the preregistration. First, the two geographical variables elevation(average elevation of each district, in meters) and dist (distance from Kabul, in kilometers) were rescaled. The former was rescaled to elevationk (in kilometers) and the latter was rescaled to dist_le5 (in 105 meters). Second, in the original study, the clustering variable used for estimating cluster-robust standard errors was regional commands (N=6). In the present study, the clustering variable has been changed to "province" (N=34). Please refer to "General Discussion" for detailed justifications.; The deviations are the same as those reported in Trace #3: First, the two geographical variables elevation (in meters) and dist (in km) were rescaled to elevationk (in km) and dist_le5 (in 105meters). Second, the clustering variable used for cluster-robust standard errors has been changed to "province". Please refer to "General Discussion" for detailed justifications.
kk47	7qnxh	Deviations from preregistration: None
618k	x9j47	Deviations from preregistration: It was originally planned to fully randomize across the three experimental conditions (NoBrand vs. BrandA vs. BrandB). However, due to the randomization error, one condition was collected first (BrandB), and the other conditions were randomized subsequently. However, no data analysis was conducted before the data collection was completed. (subsequent post-hoc analysis) It has not been pre-registered, but the post-hoc test with Tukey multiple comparison corrections result showed that the benefit - symbolic difference score was significantly different for the symbolic product category (t(923) = 3.78, p = .0002), while that for the utilitarian product category was not significantly different (t(923) = -1.839, p = .0662). This pattern was consistent with the original study.
6z3o2	cnbt5	Deviations from the preregistration: We made a small (but influential) wording mistake in the inferential criteria concerning the effect size. In the pre-registration we wrote: "Its effect size should however not be above -0.1, indicating that cultural possessions marginally negatively moderate the BFLPE." But this is not correct. Instead, its effect size should not be below -0.1, indicating that cultural possessions marginally negatively moderate the BFLPE. We have corrected this mistake and specified the correct inferential criteria here above.
4zz0	q7yc5	Deviations from the preregistration: There are no deviations from the preregistration.
g9mm	gsw9a	There were some deviations from the pre-registration. We had estimated that we would need to recruit 312 participants to meet the required sample size of 302. However, after excluding data based on pre-registered exclusion criteria, the required sample size was not met. We recruited an additional 30participants, taking the total number of recruited participants to n=342. Another deviation from the pre-registration occurred after an unexpectedly large proportion of participants (58%) reported a goal that was not 'interpersonal' when they had been instructed to provide an interpersonal goal. Therefore, we changed the instructions to include a definition of 'interpersonal' when participants were asked to provide an interpersonal goal(as can be seen in "SCORE -Kappes_JournExpSocPsych_2012_9J1 -Direct Replication –g9mm–Inquisit script amended.iqx"). Thus, the instructions changed from "Please name your most important interpersonal goal" to "An interpersonal goal is a goal relating to relationships or communication between people. Please name your most important interpersonal goal."This meant that the additional 30participants that we recruited received the altered instructions. There was a minor deviation from the pre-registered data file as there were five additional variables included to code the data for those that met the exclusion criteria of having <2 trials for each trial type, used to calculate the mean for each trial type.First, the variables, "IntAssocObs" and "IntAccessObs" were included where participants who met the criteria for exclusion were coded as '1' and participants who did not were coded as '0'. The variable "filter_$" coded participants who were selected for further analysis of the main study analyses. A new dataset of the data which met the inclusion criteria, "SCORE -Kappes_JournExpSocPsych_2012_9J1 -Direct Replication –g9mm–focal and additional analytic dataset cleaned.sav", was produced for analyses where interpersonal obstacle-behaviour associative strength was the dependent variable. The variables, "HealthAssocObs" and"HealthAccessObs" were created for participants who met the criteria for exclusion for the health obstacle-behaviour trials and interpersonal obstacle-health behaviour trials.These variables were coded as '1' for participants who met the exclusion criteria and as'0'for participants who did not. Two new datasets of the data which met the inclusion criteria, "SCORE -Kappes_JournExpSocPsych_2012_9J1 -Direct Replication –g9mm–H8analytic dataset.sav"and "SCORE -Kappes_JournExpSocPsych_2012_9J1 -Direct Replication –g9mm–H9 analytic dataset.sav", were produced for analyses where health obstacle-behaviour associative strength and the associative strength of health goal obstacle and interpersonal behaviour were the dependent variables. A separate analysis script was created to compute the remaining required variables for the H8 and H9 analyses, "SCORE -Kappes_JournExpSocPsych_2012_9J1 -Direct Replication –g9mm–health cleaning script.sps" A minor amendment was made to the focal claim analysis script to include a command to always probe interactions for thePROCESS models, as the default is to probe interactions only if p<.10 for the interaction. A deviation was made from the pre-registered analysis plan for the additional analyses that were conducted.The pre-registered analysis plan stated that H3 and H4 would be tested in a moderation analysis where condition (variable named 'Cond2' with mental contrasting as the reference category) would be entered as the independent variable. It was noted and corrected that replicating the proposed effect for H3 could be achieved in the analysis that was being conducted to test H6, where a moderation analysis was being conducted with condition (variable named 'Cond3' with reverse contrasting as the reference category) as the independent variable. Replicating the proposed effect for H4 could be achieved in analysis that was being conducted to test H* and H7 where a moderation analysis was being conducted with condition (variable named 'Cond1' with irrelevant contrasting as the reference category). These amendments have been noted in the analysis script.
10g2	2cpf8	Deviations from the Preregistration There were no deviations from the preregistration.
yzz0	rg5vt	Deviations from the Preregistration None.
249w6	x46bu	Deviations from the preregistration: There are no deviations from the preregistration in terms of estimation procedure. However, our team included another estimation results which only use pre-Covid period data (from Jan 2008 to Dec 2019) as robustness check. By imposing 'pre-Covid' rule, the last 12 observations (from Jan 2020 to Dec 2020) are excluded.
ykk0	uvgp5	Deviations from the Preregistration There were no deviations from the preregistration.
2w9k2	dh3ae	Deviations from the preregistration: I discovered a typo on the instructions for R=9 on page 15, where it should read "majority of 9" instead of "majority of 5". I corrected the typo when printing.

m7m3	5qet4	Deviations from the Preregistration A few minor changes to the data analysis script were introduced: 1. packages ‘meta’ and ‘metafor’ are required but were not listed at the beginning of the script. It is corrected in the updated version. 8 2. in line 115 of the updated script the order of the variables is changed to correct an error in the preregistered script (number of studies with ‘percent female’ and number of studies with ‘mean age’ variables had wrong order). 3. in line 55 of the updated script ‘percent female’ variable is converted to numeric to match variable types and successfully merge the original and replication datasets. I also planned to start the initial search in databases in July 2020 but actually started on August 11th 2020.
6547	s2hez	Deviations from the preregistration: There are no deviations from the preregistration.
21k52	r6edv	Deviations from the preregistration: No deviation from pre-registration.
g241	j3u2s	Deviations from the Preregistration There were no deviations from the preregistration.
y050	mevd9	Deviations from Preregistration: In order to achieve a desired sample size of 30 observations, we obtained supple- mentary data for Trinidad and Tobago that was not originally included in the preregistration. Specifically, we computed the 2010 trade share of GDP (112.2845) using data from https://oec.world/en/profile/country/tto and https://data.worldbank.org/indicator/NY.GDP.MKTP.KD?locations=TT, and we computed the average ‘General gov- ernment final consumption expenditure’ as a share of total GDP between 2010 and 2014 (10.1) using data from https://unstats.un.org/unsd/snaama/CountryProfile. Also, the Polity5 series was used instead of the Polity IV version for democracy scores.
93k7	h7jfv	Deviations from the preregistration I thank a reviewer for recommending testing the model fit for varying numbers of latent classes using Bayern Information Criterion. I include this in the analysis code; however, each result suggests that 4 classes fit the data better than 3 classes. The original authors show similar results in their online supplement (found here), but decide to use 3 classes due to the results of Lo- Mendell-Rubin likelihood ratio tests. That test is currently unavailable in Stata. The inferential criteria in the preregistration suggests that the replication relies on the result of a one-tailed test (given the post-secularists will have significantly lower values than the traditionalists); however, the SCORE project team requires all results be shown as a two-tailed test therefore this document only provides the results of the two-tailed test.
2wgk2	957xr	There were deviations from the preregistration.
6m33m	qfh2r	Deviations from the preregistration: None.
2y2g	5fvdy	Deviations from the preregistration: We have followed the plan stipulated in the pre- registration
6m4k	kms49	Deviations from preregistration: There were no deviations from the pre-registered design and procedure. The analysis plan was updated after data collection to reflect minor unexpected features of the data (i.e., to remove the Prolific ID and screen out incomplete responses). Please see the “AnalysisScriptV2.Rmd” for details. Although not a deviation, there was one mistake in the pre-registration document. Whereas the hypotheses were correctly specified, the details in the inference criteria section were incorrectly worded. More specifically, the correct wording should say (changes in bold): “For this study, we will also use the same criterion to test our additional hypothesis (H1). Our three-level dummy for the condition will generate two coefficients, one for each hypothesis. The test of H* corresponds to the coefficient on the dummy capturing the effect of the price being consistent at low price (relative to the reference point) and the H1 is tested with the remaining coefficient on the dummy capturing the effect of price being consistent at the high price (relative to the reference point). Both hypotheses use the same criteria (alpha = .05, two-tailed).”
288g2	t3py6	We did not deviate from the pre-registration.
k5z	c2hry	Deviations from preregistration: The only deviation from the pre-registration concerned the exclusion criteria. With the original aim of utilising only high-quality data, we pre-registered that we would exclude couples who gave inconsistent responses to the question “Do you currently live with your partner? (If you live with them some of the time, please select "yes" if it is the majority of the time)”. Members of N=5 couples gave inconsistent responses to this question. However, we decided not to exclude these participants. This was because we wanted to maximise the sample size given that we recruited fewer couples than anticipated, and the fact that in hindsight, the question is open to interpretation given that members of a couple may legitimately disagree on what constitutes “the majority of the time”. Two of these couples were also inconsistent on the other exclusion criteria question of “length of relationship” but were retained because the discrepancy in length was minimal (<1 year).
8zz7	xt4j7	Deviations from the Preregistration There were no deviations from the preregistration.
0y68	6r5de	Deviations from Preregistration There were no deviations from preregistration.
1y02	dscfg	Deviations from preregistration: There were no deviations from the preregistration.
23w8w	39upt	Deviations from the preregistration: There are no deviations from the preregistration in terms of estimation procedure.
4142	8hnxg	Deviations from the Preregistration There is no deviation from the preregistration.
15yz	98j2m	Deviations from preregistration: Data were collected from a total of 267 participants due to their availability, but the first 132 are included in this report in alignment with the project’s preregistration. Additionally, all participants were recruited from psychology courses for extra credit. Thus, paid recruitment and related study materials were not utilized.
kzyz	pex6u	Deviations from the Preregistration There were no deviations from the preregistration.
z161	nftez	There were no deviations from the preregistration including in recruitment, design and analysis.
0056	e9dn4	Deviations from the preregistration: There are no deviations from the preregistration.
658y7	skwev	Deviations from the preregistration: The models were saturated and anova didn’t produce p-values for models comparison so estimated marginal means were used to compare intercepts instead and code added to the analysis file. The updated file with the analysis code was added to the OSF as Version ID 2.
gz51	7efsw	There are no substantive changes made from the pre-registration
m7g7	9km7b	Deviations from preregistration There was no deviation from the preregistration.

3g4k	s5uky	Deviations from the preregistration: There are no deviations from the preregistration.
6zy82	675hz	Deviations from the preregistration: There are no deviations from the preregistration.
8z2g	w9acu	Deviations from the preregistration No deviations from the preregistration occurred.
yk91	q4e35	Deviations from preregistration: In Stage 1 we preregistered to collect data from 218 but we sent the online survey to a bigger pool. We had 346 correct responses when we closed down our first data collection period. In Stage two we collected only 139 participants to sum up to the preregistered total participants.
mkz7	ceaq3	Deviations from preregistration: it was noticed that the code makes sense when the variable replication_data_mun (in the 94 line of the code, and following lines) is changed to replication_data. This does not change the results of the focal analysis.
38	zhdtn	Deviations from the Preregistration None
1642	78xj3	Deviations from the Preregistration: The study was run as pre-registered, with no notable differences.
23m12	34qhg	Deviations from the preregistration: None
99m7	f93t6	Deviations from preregistration: No changes were made to the preregistered script/syntax.
1554	ab762	Deviations from the Preregistration The target sample size in the preregistration was 460 participants with the aim to get 396 for the analytic sample. We had around 1,000 participants start the study, but some dropped out without completing the task and of those who finished many were not eligible to make it into the analytic sample because of the exclusion criteria outlined in the preregistration (primarily not watching the video, not clicking the button, and not being able to answer questions about the video). Secondly, we (the replication team) were working under the assumption that we needed to run a balanced two-way ANOVA and thus had to get an equal number of observations in each cell. Due to our misunderstanding, we kept running batches of participants until we had at least 99 4 observations in each cell (2 Nationalities x 2 Conditions) to meet the analytic sample target of 396. However, in retrospect, we could have stopped sooner when we reached 396 even if some cells were below 99. Ultimately, running (a) an unbalanced ANOVA with 396 participants (which discards observations in chronological order after 396th), or (b) an unbalanced ANOVA with our final analytic sample of 448 participants (which includes all observations that satisfied the inclusion criteria), or (c) a balanced ANOVA with 396 participants (which discards observations in excess of 99 from each cell), confirms the focal hypothesis. Lastly, during preregistration we took the editors' advice and removed controls from the video so that participants were not able to skip around. As such, we did not ask participants if they skipped around the video. However, we forgot to remove that from the "exclusion criteria" in the preregistration. In other words, participants were not able to skip around the video, they were not asked about skipping around the video, and so that exclusion criteria was not in force.
g5m	9bme4	Deviations from preregistration: Several differences from the preregistration are noted. First, although we retain the specified multilevel mixed-effects logistic regression model as a supplemental analysis, we elevate the population-average logistic regression model to be primary for purposes of testing the SCORE claim (Szmaragd et al., 2013). We also do not use sampling weights in the analyses. Both of these changes were made after consulting about the replication with the original author Bersani. We also revised the analytic script to correct minor errors and to export results in tabular form. This revised script is available on the OSF project page (https://osf.io/wmnrx/).
944y	bc2v6	Deviations from the Preregistration There were no deviations from the preregistration.
2w5k2	3ykw5	Deviations from the preregistration: There are no deviations from the preregistration.
g79m	vspzr	Deviations from the Preregistration There were no deviations from the preregistration.
k6y7	nuzh2	Deviations from the Preregistration There were no deviations from the preregistration.
95my	uc5zb	Deviations from the preregistration: None.
y496	9t4nh	Deviations from the Preregistration There were no deviations from the preregistration.
9ky	axph8	Deviations from preregistration. The preregistration stated that the initial recruitment target for stage 1 data collection was 31 participants to achieve the planned sample size of 26. Based on the exclusion criteria stated on the preregistration, 15 participants were excluded due to not completing the entire experiment, equipment malfunction, and/or their data not showing an expected effect of monotonically better target identification at longer target durations. Additionally, 1 participant was excluded due to reporting having a neurological diagnosis, and 6 participants were excluded due to their data showing negative t0 values in either or both conditions, indicating a perceptual threshold of less than 0 milliseconds as calculated by the original equation with the given parameters. To not change the original equation or parameters, data were collected from 48 participants to achieve a sample of 26 participants for data analysis.
69812	nr64y	Deviations from the preregistration: The replication study did not deviate from the design brought forth in the preregistration form.
8z81	7qjzp	There are no substantive changes made from the pre-registration.
7g66	wmqfu	Deviations from the Preregistration We did not deviate from the preregistration.
69136	bfs6p	Deviations from the preregistration: None
9977	pbw2f	Deviation from preregistration I moved the study online (as opposed to in person) due to the COVID-19 Crisis (prohibiting in person research for the foreseeable future).
28yg6	7q5d4	Deviations from the preregistration: There are no deviations from the preregistration.
329k	kcegt	Deviations from the preregistration: No deviations.

yk16	ybzev	During data collection of Stage 1 we observed a great difference in proportion of complete responses in the Politics group and in the Non-Politics group. The Non-Politics group was already completed and we were observing a rate of one participant in the Politics group for each ~21 screened out due to quotas full in the Non-Politics group. This was considerably slowing data collection and was also problematic to the survey company. Upon consultation with the SCORE/DARPA coordination, in early November 2020 we decided to relax the inclusion criteria for the Politics group. This change was made prior to the data analysis of Stage 1. Thus, it was not influenced by the knowledge on whether we could or not replicate the claim we are attempting to replicate. An addendum to the preregistration was included at OSF with a detailed explanation of the change (https://osf.io/kvjeh). In sum, differently than stated in the preregistration, we decided to include in the Politics group individuals who select Politics or History or Public Administration or Government or Sociology as topics important for their studies/work. People who select any other topic compose the Non-Politics group. These topics were chosen because they require a certain knowledge on Politics. The same criteria was implemented in data collection of Phase 2. Complete responses in Stage 1 and 2 were recoded according to this new criteria. Due to these changes in criteria, categorization into either the Politics or the Non-Politics group had to be conducted manually (as opposed to the pre-computed categorization implemented in Qualtrics). As this was feasible only by downloading the raw data as a .sav instead of a .csv file, minor changes in the recoding of variables had to be added to the data analysis script. Furthermore, we added code to conduct contrasts of the estimated marginal means in the ANCOVA (H2, H3a-c) in order to correct for multiple comparisons. The following changes were made to the pre-registed data analysis script...
23yw2	3d97y	Deviations from the preregistration: There are no deviations from the preregistration.
651m6	gn3vj	Deviations from the preregistration: None.
247z3	5r28d	We tweaked the code based on Lodder et al. (2019) so counters will be displayed correctly. Furthermore, note that although Vadillo et al. (2016) work with Cohen´s d as an effect size, Hedges g as an effect size was used in this attempt as it was used in Lodder et al. (2019).; We conducted a sensitivity analysis with corrected code.; We conducted a sensitivity analysis with improved code based on Lodder et al. (2019).; There was an error in inferential criteria for this claim as the p-value should be more than .05, not less. Although the claim started that... “The last two rows in Table 1...”, a more specific analysis of the claim and focus only on the last line (Table 2 in Vohs, 2015) provided non-significant findings. Thus, it should be stated that fixed-effect and random-effects effect size estimates in Table 2 will NOT be statistically significant. This is in line with the present results.
288ok	syq4e	Deviations from the preregistration: None.
65o96	e9j6y	Deviations from the preregistration: None.
308	ftv7q	As we had to collect data online due to the pandemic, data collection deviated from the pre-registration in one aspect. At the end of the study, participants were not asked whether they were aware of the purpose of the study. The data analyses also deviated from the pre-registered script (Kang_308_Syntax_en.sps) in a few minor ways. First, we created two different scripts based on the pre-registered script in order to run the claim analysis script (Claim analysis.sps) separately from the full data analysis script (Full script.sps). Second, the code was slightly adjusted to account for changing document names to make the final report more homogenous and comprehensible.
3053	b79mc	Two deviations must be reported. First, we describe an error in the analytic script used to reanalyze the focal claim and the steps taken to produce abetter replication attempt.Second, we describe how the ordering in which missing data were removed was revised to better match what was described in the original article.Error in analytic script. On page 405 of the original article (Seaton et al., 2010), the original authors specified the fixed components of each multilevel regression model. Specifically, they wrote that “…the fixed components were individual ability (both linear and quadratic), the specific moderator, school-average ability, and the cross-product of the moderator and school-average ability.All models tested had a three-level structure: Individual students were at Level 1, schools were at Level 2, and country was at Level 3.” The Data Analyst followed these and other procedures outlined in the Methods section of the original article to produce the multilevel regression models needed to replicate the focal analysis.Unfortunately, the original authors did not specify the random components of each multilevel model and, to the best of our knowledge, the Methods section of the original article indicates that random effects for school and country only needed to be specified.However, an examination of Table 3 (see p. 411) indicates that the random effects components included Level 3 (country intercept, linear ability, quadratic ability, school-average ability, and moderator), Level 2 (school intercept, linear ability, quadratic ability, and moderator), and Level 1 (individual intercept) variables.Thus, the original models(i.e., the ones used to produce the original replication result) were not specified correctly –some of the random effects were not included. This discrepancy was brought to the SCORE team’s attention on Tuesday, October 27, 2020. On Wednesday, November18, 2020, the Data Analyst was given permission by the SCORE team to revise the analytic script to include the missing random effects and update the replication analysis result. Both the SCORE team and the Data Analyst agree that the revised analytic script offers a better test of the focal analysis claim and, thus, should be used as evidence for the SCORE project, rather than the preregistered analysis. Taken together, the Data Analyst revised the analytic script multilevel regression models to match the output (i.e., fixed and random effects) reported in Table 3 of the original article(see https://osf.io/qrcpe/?view_only=f8c113e4a21b45bca06c091a5db3a411). However, it is not clear why a Level 1 variable (i.e., the individual student) was included in the random effects component of the model because there is only one observation per individual in the dataset. It would make sense to include a random effect for the individual if multiple observations were recorded for each student, but an examination of the replication dataset indicates that only one observation per student was recorded. As such, we are not sure how a random effect can be captured for the individual intercept when there is only one observation per “unit” at Level 1.Thus, the analytic script was revised to include all random effects reported in Table 3 of the original article(except for the random effect at Level 1 because is only one observation per student). Missing data. In the original analytic script, observations were removed in the following order.(1) Observations nested in schools with 10 or fewer students were removed(2)Observations with missing data in the focal variables were removed(i.e., math self-concept [DV], plausible values 1-5 [IVs], and memorization [moderator]However, a closer inspection of the Methods section indicates that the ordering in which observations are removed should have been reversed (see page 403of the original article; https://osf.io/qrcpe/?view_only=f8c113e4a21b45bca06c091a5db3a411).As such, in the revised analytic script, observations are removed in the following order:(1)Observations with missing data in the focal variables were removed(i.e., math self-concept [DV], plausible values 1-5 [IVs], and memorization [moderator](2)Observations nested in schools with 10 or fewer students were removed
7965	3quzn	Deviations from Preregistration There were no substantive deviations from the pre-registration. Two small deviations are described below in the interest of full transparency. 1. On page 20 of the pre-registration, an error was made. We wrote: “The regression coefficient and p value associated with the link between MAv_FAC [Mastery Avoidance] and REH_FAC [Rehearsal] are the statistics of interest.” This was a typo. This statement should have read “... MAp_FAC [Mastery Approach] and REH_FAC [Rehearsal] are the statistics of interest” (bracketed words added for clarification). As reflected in our claim description on page seven of the pre-registration and the analysis script, the relationship between Mastery Approach and Rehearsal was the claim of interest. 2. The Rscript was modified to load and clean the data gathered from Qualtrics and also to output more summary statistics from the model. The focal model, itself, was not changed.
yyy1	fzut6	There were no deviations from the preregistration including in recruitment, design and analysis.
2y582	4w3dy	There were no deviations from the preregistration.
2kwg2	d5ezq	Deviations from the preregistration: There are no deviations from the preregistration.

6mz92	zr7dw	A few deviations from the pre-registration occurred. The pre-registration did have a typo where the ideological sophistication scale was referred to as the political sophistication scale. All answers were coded by a blind coder. In looking over the coding, the blind coder was relatively flexible in terms of the coding. For example, participants were not required to spell names correctly in order to get questions correct. Therefore, a different coder may come up with slightly different coding structure. There is also an error in the pre-registration where the answer for “Generally speaking, is raising taxes more economically liberal or more economically conservative?” is listed as conservative. This was coded correctly by the coder. In addition, for “What is the official name of the major piece of healthcare legislation signed into law by Barack Obama in March of 2010?”, only Obamacare was not accepted as Obamacare was not the official name
mk67	579wh	Deviations from Preregistration: The replication differs from the preregistration in minor ways. First, we had proposed to alter the scaling of two control variables (electrification, elevation) to aid interpretation of results, but we chose to keep the scaling the same as the original in the final analyses. Second, we revised the analytic script to export results in tabular and graphical form. This revised script is available on the OSF project page (Weidmann_Data_Analysis_Final.do).
65gm6	b2xma	Deviations from the preregistration: The 02_anlysis.R code was changed slightly to (a) include “library(dplyr)” and (b) to uncomment out the portions of the code required to run on the full dataset.
6ow46	r42en	Deviations from preregistration We planned to collect data online from the University of Exeter FEELE participant pool and or the online survey platform Prolific. However, because of high dropout rates online, we collected the bulk of the data from FEELE participants in the lab. In line with lab rules, participants who turned up to an in-person session but could not participate were paid £5. To reduce the chance of dropouts when data was collected online, we created a Qualtrics survey that presented the consent form and then asked participants to indicate that they understand that (1) the study takes 25 minutes and are able to commit to completing it and (2) if they drop out before the session is complete the data from the session will not be usable, costing the experimenter time and money.
41k2	gzwhe	Deviations from the Preregistration 1. While in the preregistration, “White” is included as a control variable in the model, in the replication analysis, it corresponds to the reference group for race.
2wyw2	59vgy	Deviations from the preregistration: None.
m63	bnauj	Deviations from preregistration: There was a small variable naming bug in the preregistered R script. The baseline/dual-task factor expected values of baseline and dual in the script, but the experiment script generated values of baseline and load. These naming errors were fixed. We did not preregister the part of the analysis script necessary to generate a p-value for the final result described in the other results section in the event that the effect was significant instead of null. This script was added in for completeness.
2ggz2	ahgfx	In addition, no deviations from the preregistration occurred.
7g95	4bckg	Deviations from preregistration: Our replication project involves no deviations from the associated preregistration.
gz2m	b9rcg	There was a major deviation from the preregistration including in recruitment, design and analysis due to COVID-19. Due to COVID-19, this study was conducted online on Amazon.com Mechanical Turk using photographs rather than using a real table in the lab.
9kzy	76y35	Deviations from preregistration In order to have 461 participants completing both T1 and T2, we preregistered that we will sample 551 participants at T1. We ended up sampling 606 participants at T1, and 517 students took part at T2. We were able to match data from both timepoints for 443 of them. The main reasons for this were that some participants made errors when self-generating the code we used for data matching (this was done to ensure participants' anonymity) or ignored the instructions for code generation. Another deviation for preregistration plan was that due to COVID-19 pandemic all participants completed the questionnaires remotely, whereas we initially planned to have some of them complete them during class time, which likely would have reduced both the attrition rate and the number of incorrectly generated codes.
om7	acdve	Deviations from preregistration: There were two deviations from the preregistration that should be mentioned. First, the preregistration stated that the AOWS measure would be adjusted by subtracting 112 from the sum of the items and then the log would be taken. Issues taking the log led to conversations with SCORE coordinators and SCORE statistical consultants. The decision, based on what was done in the original paper, was to take the log of the absolute value of AOWS after subtracting 112 (+1, which was also done in the original paper). The second deviation is the sample size. The initial pre-registered sample size (n = 169) was communicated in error, and the final sample size based on a power analysis run by COS was 183. Analyses were done after the originally preregistered stopping rule with (n =171), which revealed a significant improvement to model fit with the addition of WICS (i.e., the focal hypothesis being replicated was supported). Additional responses were collected at the request of COS after results were known in order to meet target power requirements for the focal analysis. That is, sampling beyond the pre-registered sample size of 169 occurred to meet the true targeted sample size of 183. Once 183 participants had been recruited, a review of the self-reported age variable in the survey revealed that 18 participants reported being under the age of 41, despite the Cloud Research panel feature that was used to exclude participants under the age of 41. These 18 participants were re-sampled. However, re-sampling 18 participants revealed that 4 out of the 18 participants self-reported being under the age of 41 (despite again using the Cloud Research panel feature). These 4 participants were re-sampled and all 4 re-sampled participants were at least 41 years old. In total, 22 participants were re-sampled to get to the analytic sample size of 183.
24m16	c6a8e	Deviations from the preregistration: None
165m6	4gd23	Deviations from the preregistration: To ascertain a higher degree of belonging to a specific nationality, further screening tools were used during participant recruitment, requiring that participants were also born and living in the country of interest (this last condition was not applicable to the Chinese and Indian subsamples due to the limited participant pool available through the recruitment platform).
786	jqrde	Deviations from the Preregistration There were no deviations from the preregistration.
2k5g2	ck8aq	Deviations from the preregistration: None
6g38	hzg3j	Deviations from preregistration There were three deviations from the procedure described in the preregistration document, to which we refer in the following. (1) We had originally planned to collect data from 43 participants in stage 1 and data from another 20 participants in stage 2, if necessary. For logistical reasons, all data were collec- ted in one go. Since the data are time stamped, we decided to use the data of the first 43 participants for the stage 1 analyses and then add the data from the remaining participants for the stage 2 analyses, if necessary. This decision was made in consultation with the SCORE coordina- tors. (2) We excluded the data from one participant (i.e., R_e4KcPgvKpQpamWd) prior to the analyses, because this participant wrote random letters in response to some of the questi- ons (e.g., Q45, Q52, Q64, Q78, Q92, Q99). We did not anticipate this and therefore did not describe a standard procedure for such a case in the preregistration, but decided to exclude the data from this participant neverthel- ess, because it was unclear whether they answered the rest of the questions honestly, or not. (3) We had originally planned to collect data for a pooled sample of 63 participants, but fi- nished data collection with usable data from 68 participants. When the data from the addi- tional five participants were omitted, the statistical results remained unchanged, so we de- cided to use the data from all 68 participants. This decision was made in consultation with the SCORE coordinators. In all other respects, we followed the procedure described in the preregistration document.

6mz8	fxwt7	Deviations from preregistration 1. We recruited 154 participants in the second stage instead of 153 due to a mistake when setting up the recruitment platform (Prolific.co). 2. In the preregistration, based on Figure 4 in the original paper, we interpreted that Bruner calculated the correlation using data of the unique combinations of outcomes instead of all observations. For example, if three people had a 0.3 error rate and an average number of safe choices of 5, then we collapsed these three observations into one. Using this method, the correlation is estimated with less information and each observation is not correctly weighted to reflect the number of participants with a particular combination of outcomes. Expost, we reconsider our interpretation. To avoid ambiguity, we manually imputed the original data from Bruner based on Figure 4and calculated the Pearson correlation (see folder "withDataImputedFromBrunerPaper"in the Analysis component). Our analysis confirms that our initial interpretation was incorrect, and Bruner indeed used all observations (However, we note that we did not get exactly the same results. Specifically, Bruner find a correlation of -0.4625 and we find a correlation of -0.4484). We rectify our mistake in this report. Specifically, in code used for the analysis (ReplicationFile.do), we comment out line 215 in which we used to collapse observations into unique combinations of outcomes. Everything else remains constant.
m5g9	r6ynf	Deviations from the Preregistration In the preregistration we specify that only Black U.S. citizens would be allowed to take the study. Upon deployment of the study, a member of the Scientific Advisory Committee for Project Implicit expressed concerns that limiting the study to only Black participants would negatively impact the collection of Black participants for other studies in the research pool, therefore the study requirements were adjusted to allow U.S. citizens of any race to take the study. Importantly, only data from Black participants is analyzed for the purposes of this replication.
mkk9	rzny2	Deviations from preregistration There were no deviations from the original preregistration.
42k8	j5y4t	Deviations from preregistration: 1. The results were initially reported using the preregistration script and showed no interactions between variables of interest for the original data sample. Upon further investigation it was revealed that the original script lacked log transformation for the cholesterol data in the original Japanese sample. The missing transformation (cholesterol_log = log(cholesterol)) was added to the code. 2. In the preregistration code the recoding of negative affect variables was based on positive affect variables rather than the underlying negative variable. It was corrected and all negative variables were recoded based on negative affect variables. 3. In the preregistered analysis code culture was incorrectly specified as a random effect. It was corrected by replacing culture with participant ID as a random effect. No other deviation from the preregistration occurred and the same inferential criteria as reported in the preregistration were applied upon correction. The analyses were repeated resulting in associations reported above. Updated analysis code was uploaded to the OSF.
5zg9	6gpkx	The data analyses deviated from the pre-registered script (SeunjensScript.R) in a few minor ways. First, four packages (i.e., "Hmisc", "corrplot", "multcompView", and "xfun") were required to run the final markdown script that were excluded from the original script. Second, code was added to the scripts to properly import the .csv file downloaded from Qualtrics. Third, errors in the code written to produce descriptive statistics were discovered and corrected. Fourth, correlation analyses that were included in the pre-registration text but excluded from the registered analysis script were added to the final markdown script. Code changes are explained and noted in the full R markdown script (SCORE - Seuntjens_JournPerSocPsy_2015_PNPz - Schmidt_5zg9 - R markdown full script.Rmd) and its output (SCORE - Seuntjens_JournPerSocPsy_2015_PNP - Schmidt_5zg9 - R markdown full Output.docx).
234w6	tv4se	The Irish team aimed to collect 200-236 participants. With a projected survey completion rate of 47%, based on the original study, this meant we would send out 502 questionnaires. However, due to a low response rate we needed to contact more administrators. In total we emailed 4,614 preschool administrators. This resulted in 422 participants signing up for the study, of which 392 completed Time 1 measures, 293 completed Time 2 measures and 216 completed Time 3 measures. However, some participants were delayed in their completion of the surveys (e.g., completed Time 1 on Friday, then completed time 2 the following Sunday – meaning 9 days later instead of 2 days later). As such we created a "wave_matched" variable to identify which participants completed all three timepoints and in the correct order. This resulted in a final valid sample of 134 people. We have put the code online for this, but also for wave unmatched. To deanonymize email addresses that participants used to take part in the study we hashed them using SHA256 algorithm as part of the deidentifyr package in order to preserve anonymity and link the responses across the three waves. We also removed participants who did not complete all three waves.
23g12	gvayj	Deviations from the preregistration: None.
9k2y	fpku5	Deviations from the Preregistration There were no deviations from the preregistration.
y2312	hefyp	Deviations from the preregistration: None
23w12	xyrv6	Deviations from the preregistration: In our original OSF preregistration, a sample size of 671 was specified as the target given a power analysis by SCORE. Once the study reached the stage of data collection, the number of available participants on the Prolific platform meeting inclusion criteria (men with U.S. military experience age 18 or older) had decreased significantly from the originally estimated available sample. Thus, we were unable to obtain the desired sample size, instead reaching a final sample of 215 after removing participants for failing to meet inclusion criteria, missing data, or failing manipulation checks. As a result, we also did not run the structural models including all items from all measures as originally proposed due to concerns that models with such large numbers of parameters would be severely underpowered. We thus used the same parceling procedure that Popper & Amit applied in the original study, creating three parcels per measure, and running the structural models with the parceled data only to maximize power. Given the difference between the desired sample size for adequate power and the final sample size, our results should be interpreted with caution. In the original R code provided in our pre-registration, we did not include the code for running scale reliabilities, so this code was added in the final version of the R code. The names of data files changed in the final version of the R code to reflect the names of our data files. We also added code to test a seventh SEM model that was not pictured in the original Popper & Amit study, but was verbally described as the sixth model tested (described on page 757 of the original manuscript). "Model 6" in Popper & Amit (2009) is actually a seventh and final model tested, though it is labeled as "Model 6" in the paper (p.761). We also added a line of code for each model that allowed us to create visual representations of each structural model. No other deviations to report.
318z	xe3d2	Deviations from the Preregistration The methodology of this replication varied from the preregistration due to a miscommunication about the sample size at the stage of the replication. The replication team collected an initial sample with the understanding that the required sample size of 44 (initial sample) or 96 (full sample) was to be distributed across three participant groups. Upon initial data collection and 3 analysis, it was determined that these sample sizes referred to participants in the visual-only group. An amendment to the preregistration was filed (osf.io/gm5xr), and additional data was collected in the visual-only condition until the required sample size was collected. In the interest of transparency, the data collected in the other conditions is still on the OSF page for this project, and the files based on the new target and data collection was updated. Other than this recalculation based on power, the methodology of this replication did not deviate from the preregistration. We do report an additional Bayesian one sample t-test, as we wished to check whether there was evidence for the null hypothesis (i.e. responding was at chance levels).
556	pdurj	Deviations from preregistration There were no deviations from the original preregistration.

1yz2	fr6g8	Deviations from preregistration. During conduction of the replication, some diversions were necessary in order to complete the study. In particular: Participants were paid in British Pounds instead of U.S. Dollars. The equivalent of $3.50 was calculated in £2.00. This deviation was necessary due to the set up of the platform which allows payments only in British Pounds. It was initially stated in the preregistration form that nationally representative samples of United Kingdom and United States resides would be recruited. Subsequently, we found out that this screening option on Prolific.co is only available to cohorts larger than 300 participants. Therefore, it was not possible to recruit participants using this logic. However, the sample obtained demonstrated good representation of gender (equal distribution) and age, and an equal number of U.S./UK nationals were recruited.
24812	jtd7n	Deviations from the preregistration: No deviations from the preregistration.
3z7z	jyp7u	Deviations from preregistration In phase 1: We ran the study as preregistered. The quartile split for the data meant that the groups we compared were not exactly equal as we expected. The split worked in a way that participants were 34 and 36 instead of 33 per group. In phase 2: We ran the study as preregistered, but more participants than expected completed the study which led to a slightly higher than expected analytic sample size. The quartile split also created unequal groups with 83 and 69 instead of two groups of 72 participants.
24316	f87d9	Deviations from the preregistration: none
y2gz6	qmbxr	Deviations from the preregistration: None.
2g79y	6rpxu	Deviations from the preregistration: None.
4z88	up8xs	Deviations from preregistration. We originally proposed the analytic process without using Level-2 variables. After consultation with the SCORE coordinators, we revised it to include Level-2 variables to as close to replicate the original study as possible. Otherwise, we did not deviate from our preregistration.

Finally, in Table S6, as suggested by a reviewer, we highlight available objective indicators of each study including the original study sample size, sample size units, the original study effect size in its native units, the original study effect size in pearson's r, the replication study sample size, the replication study power estimate (using both forms reported in the main text), the replication study effect size in its native units, and the replication study effect size in pearson's r.

Table S6. Features of original and replication studies by replication attempt.

	Original study					Replication study					
Claim	Sample size	Sample size units	Effect size type	Effect size	Partial correlation	Sample size	Sample size units	Traditional power	SER power	Effect size type	Effect size	Partial correlation
0qar_single-trace	43	participants			0.38	127	participants	0.5	0.5			0.01
0wWk_single-trace	217504	observations	Odds ratio	0.64	-0.12	319719	employees	1	1			-0.11
1574_single-trace	157	group decisions	Cramer's V	0.57	0.57	198	group observation					0.7
1VeW_single-trace	14	groups			0.66	32	groups					0.08
1X9W_single-trace	62	respondants	Correlation	0.28	0.28	545	normally cycling women	1		Correlation	-0.11	-0.11
2GKO_single-trace	189	sector-years	Odds ratio	1.13	0.19	256	business sectors	1	0.63			-0.17
2lb5_single-trace	575				0.2	493	households	0.89	0.9			0.5
347d_single-trace	181	participants	Cohen's f squared	0.19	-0.4	105	participants	0.89		Cohen's f squared	0.13	-0.34
3aPw_single-trace	45434	person-year records				161855	person-year observations	1	0.95	Hazard ratio	1.93	
4q0L_single-trace	350	state-year observations			0.22	150	state years	0.19	0.45			0.21
521q_single-trace	225	individuals	Odds ratio	0.44	-0.17	722	Participants	0.97	0.94			-0.01
59bm_single-trace	226	participants	Incident rate ratio	0.84	-0.13	2006	participants			Correlation	-0.09	-0.09
5Awm_single-trace	2835	elementary students in 134 schools			0.22	14940	students	1	0.99			0.11
5JE_single-trace	127	participants	Cohen's f squared	0.04	-0.19	376	participants	0.8				-0.12
5Kgq_single-trace	119	regions			-0.43	159	Regions	1	1			0.03
5KrD_single-trace	353	individuals	Cohen's f squared	0.02	-0.13	921	participants	0.83		Cohen's f squared	0	-0.01
5Q82_single-trace	146090	observations in 21 European countries across 5 years			-0.62	79242	observations	0.52	0.31			0.23
5XEE_single-trace	271	participants			-0.18	1062	participants	0.61	0.61	Cohen's f squared	0	-0.06
5Xaw_single-trace	218	students			0.13	1116	observations	0.96	0.91			0.06
7RR2_2kl87y	107	students			-0.23	556	participants			Correlation	-0.22	-0.22
7RR2_4d3878	107	students			-0.43	556	participants			Correlation	-0.27	-0.27

7RR2_5vw85q	107	students			-0.28	556	participants			Correlation	-0.24	-0.24
7RR2_89x817	107	students	(Partial) eta squared	0.06	0.25	1096	participants			(Partial) eta squared	0	0.04
7RR2_9qjzdd	107	students	(Partial) eta squared	0.04	0.2	1096	participants			(Partial) eta squared	0	0.04
7RR2_b2pwnv	107	students			-0.26	556	participants			Correlation	-0.2	-0.2
7RR2_bd3w94	107	students			-0.2	540	participants			Correlation	-0.2	-0.2
7RR2_bn72ln	107	students			-0.05	540	participants			Correlation	-0.25	-0.25
7RR2_bolw7q	107	students			-0.25	540	participants			Correlation	-0.29	-0.29
7RR2_brwop3	107	students			-0.31	540	participants			Correlation	-0.26	-0.26
7RR2_d5dy7d	107	students			0.34	540	participants			Correlation	-0.17	-0.17
7RR2_g4xnzl	107	students	(Partial) eta squared	0.04	0.2	1096	participants	1		(Partial) eta squared	0.22	0.47
7RR2_m5jnlk	107	students			-0.27	556	participants			Correlation	-0.28	-0.28
7RR2_m84pw1	107	students			-0.07	556	participants			Correlation	-0.22	-0.22
7RR2_my485o	107	students			-0.25	556	participants			Correlation	-0.26	-0.26
7RR2_n1rp5w	107	students			0.17	540	participants			Correlation	-0.22	-0.22
7RR2_odrqvl	107	students	(Partial) eta squared	0.07	0.27	1096	participants	1		(Partial) eta squared	0	0.04
7RR2_yor7v2	107	students			-0.44	556	participants			Correlation	-0.18	-0.18
7WjP_single-trace	30	participants	(Partial) eta squared	0.15	0.38	112	participants			(Partial) eta squared	0	0.01
7X54_m6qkxx	2901	respondents				2164	individuals					
7X54_single-trace	1860	individuals	Cohen's d	0.75	0.33	1438	respondents	1		Cohen's d	-0.9	0.4
88xa_single-trace	128	individuals	Cohen's f squared	0.05	0.21	448	participants	0.93		Cohen's f squared	0.04	0.21
8R9d_single-trace	516	participants	Cohen's d	2.94	-0.83	477	participants			Cohen's d	-1.58	-0.62
8w97_single-trace	23	participants	Cohen's f squared	0.65	0.63	69	participants	1		Cohen's f squared	0.03	0.16
8wZ0_single-trace	85	participants	Cohen's f squared	0.08	0.28	667	unique participants					0.02
9J1_single-trace	97	students			0.25	313	participants		0.91			-0.05
9OK1_3n2npq	848	Moldovan communities			-0.51	224	communities	0.31	0.27			-0.23
9OK1_ppwpjr	848	Moldovan communities			-0.51	192	communities	0.33	0.24			-0.26
9OK1_q7q7np	848	Moldovan communities			0.41	224	communities	0.36	0.17			0.12

9OK1_w7d7ly	848	Moldovan communities			-0.51	224	communities	0.41	0.27			-0.26
9R9X_single-trace	47	participants	Cohen's f squared	0.16	0.38	928	participants	1		Cohen's f squared	0	0.02
9ey_single-trace	194	students	(Partial) eta squared	0.19	0.43	188	participants	1		(Partial) eta squared	0.1	0.31
AOQj_single-trace	1929	young adults			0.08	1362	individuals	0.55				0.01
AYQG_6rp9on	2236	respondents			-0.04	3238	respondents	0.71	0.32			0.01
AYQG_l8rz5d	4648	respondents			-0.03	7013	respondents	0.7	0.45			0
AYQG_w7r4vl	2236	respondents			-0.05	2042	respondents	0.23	0.36			-0.05
AgO1_21zn8w	29	observations			0.43	188	pairs of players					0.6
AgO1_4k6z71	21	observations			0.49	100	pairs of players					0.56
AgO1_5o3l5z	13	observations			0.58	130	pairs of players					0.63
AgO1_9oxpdp	38	observations			0.39	218	pairs of players					0.6
AgO1_b3z4nz	34	observations			0.4	230	pairs of players					0.42
AgO1_okp7q2	21	observations			0.49	108	pairs of players					0.56
AgO1_single-trace	24	pairs	Cohen's f squared	0.72	0.65	178	observations			Correlation	0.67	0.67
AgO1_ywl5vz	9	observations			0.65	30	pairs of players					0.62
AqDO_single-trace	1064	survey respondents			-0.06	673	Respondents	0.54	0.23			-0.09
AqDO_ywjwq1	4209	respondents			0.04	1930	respondents	0.35	0.3			0.03
AvOr_b7dlq4	208	participants	Cramer's V	-0.13	0.13	1245	participants	0.91		Cramer's V	-0.01	0.01
AvOr_grlzkw	223	participants	Cramer's V	-0.2	0.2	530	participants	0.94		Cramer's V	-0.03	0.02
AvOr_single-trace	210	participants	Cramer's V	0.24	0.24	749	participants			Cramer's V	0.19	0.19
AvWY_single-trace	48	college students	(Partial) eta squared	0.09	0.31	156	participants	0.96		(Partial) eta squared	0	0.04
BKxK_single-trace	30	test portfolios			0.36	30	test portfolios	1	0.34			0.23
BlRQ_single-trace	1020	students	Correlation	-0.39	-0.39	1342	Individuals	1		Correlation	-0.26	-0.26
Blxd_282qvz	265180	students			-0.07	62880	students					-0.3
Blxd_5821x4	265180	students			0.06	62880	students					0.15
Blxd_bl7k78	265180	students			-0.04	62880	students					-0.26
Blxd_bx4248	265180	students			-0.03	62880	students					-0.21
Blxd_single-trace	10221	schools			-0.06	5298	schools	1	1			-0.09
Br0x_jrn5ll			Cohen's d	0.43	0.4	83	observations	0.83				0.21

Br0x_pjvzxv	400	American adults	(Partial) eta squared	0.06	0.26	327	observations	0.86		(Partial) eta squared	0.04	0.21
Br0x_qn1lz8	196		Cohen's d	0.03	0.17	161	observations	0.37		Cohen's d	0.15	0.08
Br0x_single-trace	360	participants	Cohen's f squared	0.07	0.25	1020	participants			Correlation	-0.33	-0.33
Br0x_x7yjx8	198		Cohen's d	0.34	0.17	163	observations	0.37		Cohen's d	0.38	0.19
BrGp_5rpk15												
BrGp_7rxk4x												
BrGp_8njk52												
BrGp_blv93q					0.84							
BrGp_g3ov8l					0.38							
BrGp_m67j81					0.67							
BrGp_o25x3z												
BrGp_y3kqy7												
D2LY_g81lpo	7838	respondents	Relative rate ratio	0.96	-0.06	2524	individuals	0.54	0.57	Relative risk reduction	0.98	0
D2LY_mn5k2x	3062	respondents	Relative rate ratio	0.89	-0.07	1264	individuals	0.59	0.42	Relative risk reduction	0.89	-0.01
D2LY_mov8w7	7838	respondents	Relative rate ratio	0.91	-0.05	2888	individuals	0.83	0.54	Relative risk reduction	0.92	-0.02
D2LY_mrdnoj	2605	respondents	Odds ratio	0.46	-0.05	392	individuals	0.37	0.11			0.01
DEqr_single-trace	764	respondents	Cohen's f squared	0.03	-0.18	626	participants	0.93		Cohen's f squared	0	-0.05
E5qr_single-trace	62	country-years			0.25	77	country-years	0.15	0.39			-0.01
EJpm_b29xjd	397	country-years			-0.37	208	country-years	1	0.99			-0.12
EJpm_bonjlw	566	country-years			-0.43	168	country-years	1	1			-0.17
EJpm_m5y94l	566	country-years			0.45	168	country-years	1	1			0
EJpm_m8oy76	566	country-years			0.36	168	country-years	1	0.97			-0.2
EJpm_my1dxd	566	country-years			-0.44	168	country-years	1	1			0.24
EKBZ_single-trace	206	undergraduate students	Cohen's f squared	0.04	0.19	557	participants	0.93		Cohen's f squared	0.02	0.14
EQxa_single-trace	400	participants	Cramer's V	0.27	0.29	1392	responses	1		Cramer's V	0.26	0.26
EZ3x_single-trace	507	students			0.22	366	students	0.1	0.89			0.25
Eb2N_single-trace	10730	US school children			-0.06	319	5th grade children	0.05				-0.19
Ej3y_single-trace	192	participants	Cohen's f squared	0.04	0.21	952	participants	1		Cohen's f squared	0	0.02

G1Lr_single-trace	101	participants	Cohen's d	0.68	0.33	636	participants	1		Cohen's d	0.24	0.12
GNjz_single-trace	102	participants	Cohen's d	0.4	0.37	288	participants	1		Cohen's d	0.18	0.17
GXEW_single-trace	34	participants			0.35	144	participants	0.91		Cohen's f squared	0.03	0.17
J4W9_single-trace	68	University students in India			-0.28	276	participants					0.13
J7ek_single-trace	19	students	Cohen's f squared	0.44	0.55	174.4954	participant-responses (harmonic mean)			Correlation	0.54	0.54
JRpA_6ryjx7	375	districts				363	districts	0.24	0.62			-0.23
JRpA_7oy1jv	375	districts	Odds ratio	0		724	districts	0.06	0.86			-0.4
JRpA_l8j9xn	375	districts				363	districts	0.42	0.81			-0.2
JRpA_single-trace	375	districts	Odds ratio	0		724	Districts	0.06	0.86			
JRpA_zxz13n	375	districts	Odds ratio	0.61		724	districts	0.12	0.99			-0.41
JWzJ_single-trace	1142	women 15-44 years of age of all marital status	Odds ratio	0.48	-0.08	2437	female respondents	0.67	0.84			-0.03
K4ZD_single-trace	90	participants			-0.18	574	participants	0.88	0.9			-0.13
Kj9d_single-trace	20	participants	Cohen's f squared	0.88	0.68	50	participants	1		Cohen's f squared	0.02	0.13
Kybl_single-trace	9189	students			0.04	8043	observations	1	0.82			0.04
LbEB_single-trace	95	participants	Cohen's d	0.6	0.29	474	participants	1		Cohen's d	0.46	0.22
LmBx_single-trace	14	GAD patients	Cohen's f squared	0.72	0.65	63	GAD participants	1		(Partial) eta squared	0	0.02
LyLd_single-trace	359092	observations			0.04	370778	individuals			Correlation	0	0
N8pB_single-trace	79	participants	Cohen's d	0.47	0.23	364	participants	0.91		Cohen's d	0.21	0.1
Nj8V_single-trace	209	undergraduate students	Cohen's f squared	0.05	0.22	857	participants	1		(Partial) eta squared	0	0.01
NrrW_single-trace	13213	middle school and high school students in 30 schools			0.15	14057	students	1		Cramer's V	0.2	0.21
OKRy_single-trace	138	participants	Cohen's f squared	0.07	0.25	275	participants	0.9		Cohen's f squared	0.1	0.3
OYX0_2n1rrj	1200	students			0.1	13281	participants	0.93	1			0.09
OYX0_5o9y12	12012	students			0.05	13281	groups	1	1			0.09

OYX0_single-trace	1200	Spanish and Asian-language speaking ELL individuals				8526	participants	1	1			
OYX0_y5wnn9	12012	students			0.02	13281	participants	0.51	0.3			0.04
OYX0_ywd1y9	12012	students			0.07	13281	participants	1	1			0.1
OeGv_8vrpd2	29	countries			-0.51	24	country	0.99	0.33			-0.53
OeGv_bnxon4	24	countries			-0.78	26	country	0.75	0.86			-0.18
OeGv_br469p	33	countries			-0.75	28	country	1	0.97			-0.7
OeGv_gv8q78	33	countries			-0.71	28	country	0.76	0.85			-0.24
OeGv_single-trace	24	strong democracies			-0.78	30	countries	0.99	0.91			-0.21
Ovkm_single-trace	953	participants				1255	participants		1			
Ow0_single-trace	106	participants	Cohen's d	0.54	0.26	262	participants	0.9		Cohen's d	0.23	0.22
P1rY_single-trace	393	panelists			-0.18	642	participants	0.74	0.74			-0.19
PNPz_single-trace	290	MTurk workers	Correlation	0.21	0.21	510	participants	0.95		Correlation	0.33	0.33
Pb9K_single-trace	647	respondents			-0.09	2684	survey respondents	0.92	0.92			-0.01
PlDa_single-trace	18	quarters			0.94	32	Quarterly	1	1			0.42
Q1dl_single-trace	81	respondents			0.27	359	participants			Correlation	0.42	0.42
QYNq_single-trace	198	measurement			0.21	846	responses	0.79	0.91			0.14
QllV_single-trace	204	months			0.2	153	months	0.82	0.46			0.05
R8RN_single-trace	36	participants			0.39	274	participants	1				0.07
R9dv_single-trace	4409	women ever-in-union	Odds ratio	1.58	0.08	7207	individuals	1	1			0.1
RJb_single-trace	114	undergraduate students			0.19	516	participants	0.87	0.9			0.09
RYKv_3w5kdk	36962	respondents			0.11	27940	observations	0.86	0.83			0.1
RYKv_6xnk65	35963	respondents			-0.1	27940	observations	1	0.71			-0.08
RYKv_g41jx8	38102	respondents			0.26	27940	observations	1	1			0.26
RYKv_p5o238	37562	respondents			0.12	27940	observations	0.88	0.88			0.07
RYKv_single-trace	38102	respondents			0.26	36496	participants	0.41	1	Cohen's f squared	0.09	0.28
RYKv_wzn69v	38031	respondents			0.2	27940	observations	1	1			0.19

RZdL_single-trace	10080	observations			-0.11	8987	children	0.67	0.65			-0.09
RrOb_single-trace	1743	stocks			0.85	2142	Stock	1				-0.05
Rvb_single-trace	218	participants	Cohen's d	0.42	0.2	975	participants	1		Cohen's d	0.15	0.07
VB9K_single-trace	60	single women			-0.28	528	participants	0.05	1			0.01
VGAe_single-trace	94	students	Cohen's f squared	0.07	0.25	615	participants	1		Cohen's f squared	0	0.06
VOwm_single-trace	18949	firm-year observations			0.1	12344	firm-years	0.06	0.53			-0.02
Vj0p_b819vj	1502	people	Odds ratio	1.22		666	respondents	1	0.38			0.19
Vj0p_g1j462	1502	people	Odds ratio	1.22		666	respondents	0.99	0.38			0.04
Vj0p_g27k3q	1502	people	Odds ratio	1.22		666	respondents	1	0.38			0.19
Vj0p_g9dq43	1502	people	Odds ratio	1.22		666	respondents	0.99	0.38			0.04
Vj0p_mjjnpk	1502	people	Odds ratio	1		676	respondents					-0.06
Vj0p_single-trace	1502	people	Odds ratio	0.18		1666	individual respondents	1	1			-0.01
VjjX_single-trace	16	students	(Partial) eta squared	0.44	0.67	13	participants					0.43
Vpgm_single-trace	2474	individuals			-0.22	22418	participants			Correlation	-0.15	-0.15
VvZX_single-trace	270	observations			0.27	630	observations	1	1			0.36
WLkV_single-trace	124	participants	Cohen's d	-0.53	0.22	312	participants	0.86				0.08
WLpV_single-trace	24	matching group mean effort level			0.58	28	Groups of three participants	0.82		Correlation	0.52	0.4
WaYe_single-trace	284	subject-responses			-0.41	1812	observations	1	1			0
Wre_single-trace	36	students	Cohen's f squared	0.17	0.38	265	participants	1		Cohen's d	-0.11	0.05
Y8Yx_single-trace	62721	young married women			0.08	51995	women			Cramer's V	0.03	0.03
YWep_single-trace	11767	individuals			-0.75	2513	students clustered in 135 schools	1	1			-0.09
YmQR_single-trace	43576	observations			0.03	305244	Indian women aged 15 to 49, who are, or have been married	1	1			0.01
YpZZ_single-trace	81	participants	Cohen's d	1.09	0.48	454	participants	1		Cohen's d	0.35	0.17

a8jQ_single-trace	586	participants			0.09	2231	participants	0.31	0.9			0.05
aYyR_single-trace	238	soldiers			0.16	215	participants	1	0.45			0.22
aag9_single-trace	123	participants	Cohen's f squared	0.11	0.32	522	participants	1		Cohen's f squared	0.01	0.1
amYY_single-trace	106	subjects	Correlation	-0.46	-0.46	117	Participants	0.98		Correlation	-0.22	-0.22
bLe8_single-trace	1295	households			0.28	2537	households	1	0.81			-0.1
bY2A_single-trace	9493	children			-0.1	11286	kindergartners	0.63	0.56			-0.06
d24p_single-trace	109	participants	Cohen's d	0.69	0.33	358	participants	1		Cohen's d	0.36	0.18
d5v3_single-trace	499	Dutch-speaking Belgian psychology students			0.09	1270	observations					-0.01
dqKX_single-trace	201	students			0.82	978	participants	1	0.9			0.29
dxQp_single-trace	160274		Odds ratio	0.75	-0.01	110747	black students	1	0.54			-0.02
e227_single-trace	25	European countries			-0.13	27	countries (as independent regressions used for the meta-regression)	1	0.52			-0.46
e3G7_single-trace	263	participants	Cohen's f squared	0.04	0.19	379	participants	0.8		Cohen's f squared	0.03	0.16
eOQm_single-trace	77	participants	Cohen's f squared	0.09	0.28	484	participants	1		Cohen's f squared	0	0.01
eg1q_single-trace	102	students	Cohen's d	0.48	0.24	1125	Undergraduate students	1		Cohen's d	-0.14	-0.07
eg3p_single-trace	138	individuals	Odds ratio	0.28	-0.26	497	participants	0.9	0.99			-0.01
egX9_single-trace	342	daily measures			-0.28	410	responses	0.28	0.99			-0.15
exd7_single-trace	41	participants	Cohen's d	0.67	0.56	96	participants	1		Cohen's d	0.12	0.12
gRWz_single-trace	5867	participants			0.39	2438	participants	0.99	0.54			0.03
gbAY_single-trace	3	groups/cultures				5	groups/cultures	1	1			0.05
gbg4_single-trace	197	couples			0.27	924	observations	0.21	0.9			0.09
gbl9_single-trace	63	participants	Cohen's f squared	0.14	0.35	305	participants	1		Cohen's f squared	0	0.01

j2yd_single-trace	85	participants	Cramer's V	0.33		63	participants			Cramer's V	0.21	0.21
jDWN_single-trace	2347	harmonic N			0.11	729	observations	0.8	0.61			0.28
jLr_single-trace	76	participants	Cohen's f squared	0.22	0.42	113	participants	1	0.9	Cohen's f squared	0.02	-0.12
kXp8_single-trace	206	firms			0.24	174	Firms	0.05	0.69			0.15
lJ0w_single-trace	475	employees			0.31	170	employees	0.29	0.9			0.6
lkBL_single-trace	42	participants	Cohen's f squared	0.27	0.46	437	participants	1		(Partial) eta squared	0	0.05
mBL1_single-trace	20	participants			-0.99	20	participants					-0.95
mJDj_single-trace	2664	respondents	Cohen's f squared	0	0.04	37148	respondents			Correlation	-0.02	-0.02
mrZ_single-trace	88	adults	Cohen's q	0.48	0.22	750	participants	1		Cohen's q	-0.01	-0.01
mxyQ_single-trace	549	public housing authorities			-0.16	738	public housing authorities	0.98	0.9			0.02
pN7E_single-trace	7741	individuals			0.02	7101	participants					0.01
plLK_single-trace	62	participants	Cohen's f squared	0.08	0.27	403.237	participants (harmonic mean)			Correlation	0.01	0.01
pw3m_single-trace	25	countries			0.63	48	countries	0.9	0.98	Correlation	0.16	0.16
q8xv_22jnnj	17554	Country-Age			0.57	21945	country-year-ages	0.52	1			0.13
q8xv_524ll2	17554	Country-Age			-0.52	21945	country-year-ages	0.46	1			-0.07
q8xv_827ww4	17554	Country-Age			-0.59	21945	country-year-ages	0.54	1			-0.15
q8xv_l88v4q	393	Country-Year			-0.54	491	country-years	0.4	1			-0.05
q8xv_single-trace	393	country-years			-0.54	391	country-years	0.99	1			-0.54
qPxQ_single-trace	2001	participants			-0.1	3281	participants	0.94	0.94			0.1
qXX2_single-trace	27	countries			-0.01	28	countries	0.74	0.75			0.01
qYr7_1rzq75	2673	respondent judgments			-0.43	4309	respondent-valence judgment observations	1	1			-0.45
qYr7_4k31n8	7554	respondent judgments			-0.59	5704	respondent-valence judgment observations	1	1			-0.46

qYr7_9oj2rd	5659	respondent judgments			-0.41	5881	respondent-valence judgment observations	1	1			-0.4
qYr7_j6d39n	3388	respondent judgments			-0.31	6550	respondent-valence judgment observations	1	1			-0.37
qYr7_single-trace	933	respondents			-0.23	3054	participants	0.98	1			0.07
qYr7_ywrd82	5841	respondent judgments			-0.27	6038	respondent-valence judgment observations	1	1			-0.39
qgWj_9olz6z	118	societies			0.66	31	societies	0.9	0.95			0.63
qgWj_j6qlvv	73175	respondents			0.02	59584	respondents	1	0.88			0.05
qgWj_pplx84	73175	respondents			0.04	60452	respondents	1	1			0.04
qgWj_q76z8j	73175	respondents			0.02	60452	respondents	1	0.98			0.05
qgWj_single-trace	73175	respondents nested within countries			0.48	45119	participant	0.93	0.93			0.2
qgWj_x61xpy	73175	respondents			0.02	59584	respondents	0.67	0.88			0.02
rjb_single-trace	174	participants	Cohen's f squared	0.03	0.17	813	participants			Correlation	0.16	0.16
rpq_single-trace	48	students	Cohen's d	0.6	0.52	55	participants	0.91		Cohen's d	-0.37	-0.35
vGqL_single-trace	7840	Married men btw the ages of 18-65 in 32 countries	Odds ratio	1.01	0.51	6904	married men	0.78	0.64			0.49
w5dv_single-trace	173	participants	Cohen's f squared	0.07	0.26	740	participants	1		Cohen's f squared	0.01	0.08
wRvv_bz6lqp	110732	individuals nested in country-years and countries			0.13	90130	observations					0.14
wRvv_gld3y5	110732	individuals nested in country-years and countries			0.13	89939	observations					0.15

wRvv_gxvxw2	110732	individuals nested in country-years and countries			0.02	89939	observations					0.01
wRvv_m328ro	110732	individuals nested in country-years and countries			0.02	90130	observations	1	0.99			0.03
wRvv_m7zd6r	110732	individuals nested in country-years and countries			0.16	90286	observations					0.16
wRvv_single-trace	110732	respondents nested within country			0.13	17134	individual respondents	1	1			0.16
x0pA_single-trace	8808	individuals	Odds ratio	0.25	-0.1	1980	students	1	0.93			-0.19
x3KP_single-trace	85	couples			-0.32	101	partner dyads (couples in a relationship)	1	0.71			-0.56
xYbO_single-trace	368	students	Cohen's f squared	0.03	0.18	1502	Participants	1		Cohen's f squared	0	0.04
xvrb_single-trace	276	participants	Cramer's V	0.48	0.48	247	Prolific participants			Cramer's V	0.24	0.24
y3R4_single-trace	200	participants	Cohen's d	0.29	0.14	2060	participants	1		Cohen's d	-0.06	-0.03
yAPR_single-trace	3617	females			0.03	8227	females	0.93				0.01
yDyG_single-trace	189	participants			0.17	855	respondents	0.99	0.91			0.05
yJwG_single-trace	28	Danish students	Cohen's d	0.89	0.67	26	participants	0.91		Cohen's d	0.99	0.71
yQeR_single-trace	2293	business names	Cohen's d	0.42	0.14	2320	Firms	0.95		Cohen's d	0.16	0.06
yypJ_single-trace	138	participants	Cohen's f squared	0.35	-0.51	68	participants	0.9		Cohen's f squared	0.16	0.37
z4dO_single-trace	19296	individuals nested within country-years			0.16	18558	Participants	1	0.82			0.25
zK2_single-trace	1607	participants	Cohen's d	0.45	0.22	372	372 Black US citizens	0.9		Cohen's d	0.51	0.25
zN22_single-trace	40	participants	Cohen's f squared	0.12	0.33	158	participants	0.91		Cohen's f squared	0.01	0.09

zNEm_single-trace	27	students	Cohen's d	1.49	0.61	78	participants	1		Cohen's d	0.22	0.11
zb3Y_b66y26	236	institutions			0.05	222	institutions	0.73	0.8			0.05
zb3Y_bqw6wo	236	institutions			0.06	222	institutions	0.96	0.97			0.07
zb3Y_bzyz9k	236	institutions			0.12	222	institutions	1	1			0.21
zb3Y_gl7j22	236	institutions			0.12	222	institutions	1	1			0.19
zb3Y_m3dy4j	236	institutions			0.21	222	institutions	1	1			0.4
zb3Y_mp3lzw	236	institutions			0.05	222	institutions	0.35	0.83			0.07
zb3Y_mw9jx3	236	institutions			0.34	222	institutions	1	1			0.4
zb3Y_single-trace	236	institutions			0.05	225	Institutions	0.82	0.82			0.24
zekm_single-trace	229	preschool teachers in germany	Cohen's f squared	0.02	0.13	471	participants			Correlation	0.13	0.13
zlBL_9k761j	4458	participants	Hazard ratio	0.91	-0.05	4971	participants	0.67	0.76	Hazard ratio	0.96	-0.02
zlBL_bv7vrq	4458	participants	Hazard ratio	1.15	0.05	4971	participants	0.55	0.79	Hazard ratio	1.15	0.04
zlBL_g7jwq5	4458	participants	Hazard ratio	1.08	0.03	4971	participants	0.2	0.28	Hazard ratio	1.12	0.03
zlBL_md2wdj	4458	participants	Hazard ratio	0.93	-0.02	4971	participants	0.17	0.21	Hazard ratio	0.81	-0.05
zlBL_mo9wrp	4458	participants	Hazard ratio	0.87	-0.05	4971	participants	0.54	0.71	Hazard ratio	0.78	-0.07
zlBL_single-trace	4458	participants	Hazard ratio	0.91	-0.05	6126.243	participants (harmonic mean)			Correlation	0	0
zlm2_single-trace	29	undergraduates	Cohen's d	2.2	0.75	79	participants	1		Cohen's d	0.11	0.06
zmYY_g82z5k	813	respondents	Odds ratio	1.52	0.13	785	individuals	0.07	0.91			0.07
zmYY_moyz3d	430	respondents	Odds ratio	1.69	0.2	619	individuals	0.07	0.99			0.11
zmYY_single-trace	813	individuals	Odds ratio	1.52	0.13	1384	Youth respondents	0.26	0.99			0
zqwm_single-trace	7977660	students			0	49196731	students	1	1			0

Note: Cohen's w, Cramer's V, and the phi coefficient are coded as Cramer's V in this table. Effect sizes are included only if they are reported in the relevant source (original study or replication study, respectively). There is additional missingness in the partial correlation and power columns when an estimate could not be reliably produced due to incomplete reporting or non-standard methods.

Audit of replication analyses and outcomes and preparation for public release

The audit and revisions process consisted of multiple parts. Project coordinators reviewed the final reports on each OSF project to check if the outputs matched the reported outcomes in the dataset. If the report did not match or was missing outcomes, then the auditor would check the output from the code of the project. If there continued to be a discrepancy then the issue was

flagged for further review. In addition, coordinators completed checks of code to ensure code contained within each OSF project ran without error using the data that the lab provided.

We also audited the final reports and output of each replication with project team members that were not involved in conducting that replication. These auditors reviewed values in the final report and output to check if they matched the dataset. These auditors also completed specific claim reproductions. The majority of replication studies received a computational reproduction check, with priority given to replication analyses that were *not* conducted in R (since that was the language used for the reproduction checks). After that process, another group of auditors conducted a final review on the data and the output in the report and code within each OSF project. They also provided a holistic assessment of the replication analysis and provided their feedback. Project coordinators resolved any open issues themselves or in coordination with the project authors.

Following the original submission of this manuscript and before public release of the data, we conducted an audit of the OSF projects housing the replication plans, data, materials, and outcomes to verify completeness and appropriateness of shared information. This audit included an internal check for sensitive or proprietary materials and email correspondence with each lab. Each lab received a checklist that contained a step by step guide to check that their OSF project was ready to be made public. Steps within the checklist included: a check for the final report, removal of the pdf of the original paper, a check for other copyrighted materials, steps for handling data de-identification, and steps for handling IRB materials. When labs completed the checklist they emailed the coordinating team confirmation.

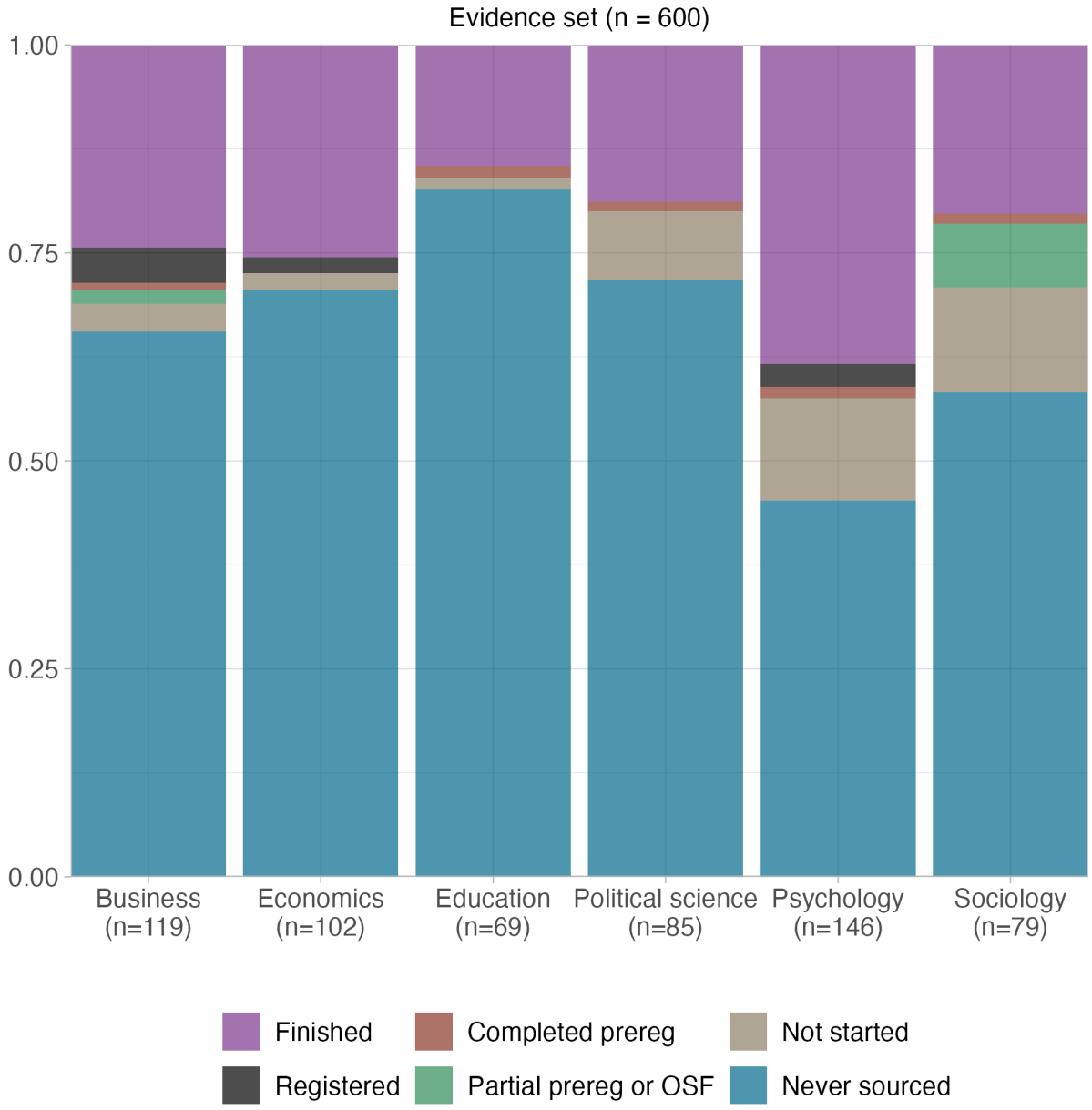
A communal tracking sheet was available to all labs as they marked completion of their checklist and observed others' completing their own. Coordinators confirmed all labs completed their checklist. This included spot-checks of the content of projects to confirm labs' reports of handling sensitive and proprietary materials properly.

For projects that were started but never completed, the coordinators followed a similar process with the individual labs and provided extra support for checking for sensitive or proprietary materials. Coordinators conducted further review for any cases in which a completed checklist was not returned.

Attrition of replication attempts selected in Phase 1

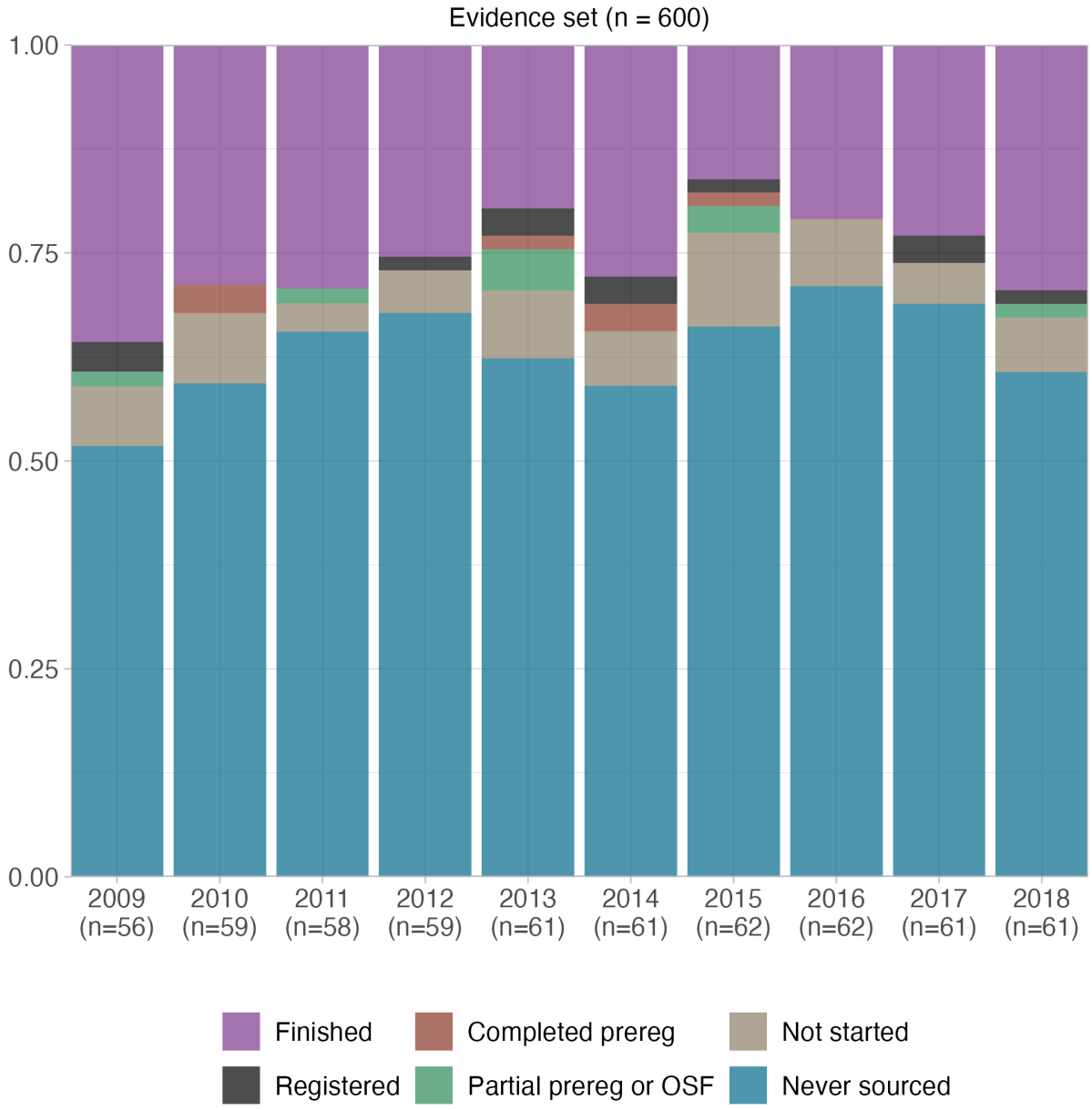
For the 600 papers made eligible for replication attempts from Phase 1, called the "Evidence set" in some documentation, we conducted additional analysis to assess attrition. Figure S7 shows the proportion of papers from these papers by discipline for which a replication was finished (purple), never attempted (blue), and not completed (other colors). Of the 63 papers for which a replication team was identified but a replication was not completed, 37 (58.7%) were never started meaning that the sourced team did not ultimately begin drafting a preregistration or worked within their OSF project for the study, 1 (1.6%) worked within their OSF project but did not begin a preregistration, 8 (12.7%) drafted a portion of the preregistration, 6 (9.5%) completed the preregistration, and 6 (9.5%%) registered the approved study. Reasons for attrition within these stages varied substantially, including a substantial portion that dropped out due to COVID, conflicting priorities in their work, preregistration designs were rejected, and insufficient time to complete the work before the program ended. Figure S8 presents the same data by year of publication, and Figure S9 presents the same data by journal.

Figure S7. Proportion of papers with a completed replication by discipline



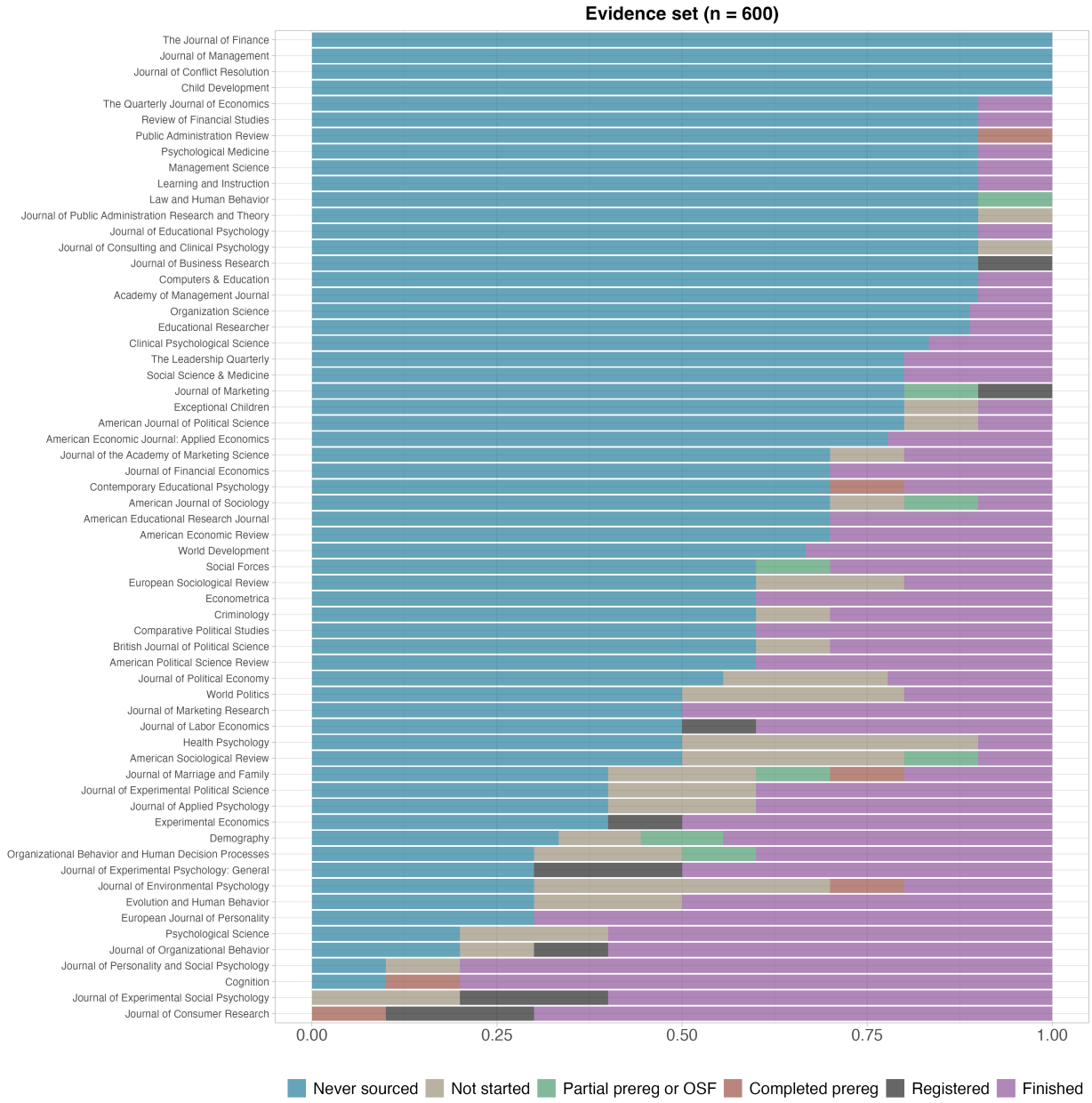
Caption: Proportion of papers by discipline for which a replication attempt was finished (purple), never attempted (blue), or for which a replication team was sourced but the replication study was not started or completed (other colors).

Figure S8. Proportion of papers with a completed replication by year



Caption: Proportion of papers by publication year for which a replication attempt was finished (purple), never attempted (blue), or for which a replication team was sourced but the replication study was not started or completed (other colors).

Figure S9. Proportion of papers with a completed replication by journal



Caption: Proportion of papers by journal for which a replication attempt was finished (purple), never attempted (blue), or for which a replication team was sourced but the replication study was not started or completed (other colors). Sample sizes per journal ranged from 5 to 10.

Non-random selection and no attrition of replications selected in Phase 2

Given the short timeline for Phase 2, we did not make a random subsample from the full set of 900 papers to be eligible for replication. Instead, we identified papers from the full set of 900 papers with extracted claims that could plausibly go through our review process and produce completed replications within the program time constraints. We engaged SCORE contributors from Phase 1 who contributed new data and secondary replications. These researchers identified 25 papers for replication (12 with new data, 13 with secondary data), all of which were

completed following review and preregistration of the study designs. Based on this approach, the 25 replicated papers from Phase 2 are likely to have more distinct characteristics from the sampling frame compared with the 139 replicated papers from Phase 1. Readers interested in probing selection effects further will benefit from distinguishing papers that were considered and replicated in Phase 1 versus Phase 2.

Excluded Cases

There are some cases in which replication outcomes became available but were excluded from the SCORE dataset for the reasons below. The hybrid replication section in the Results below reports several cases that were excluded because they combined original and independent data in the replication study, but offer interesting insights that are reported in this supporting information. Here, we summarize some infrequent cases for exclusion and explain why they occurred.

- *Replication evidence for two claims were excluded because the original claim was a negative result*, not a positive result. Positive results were an inclusion rule for claims in SCORE. Negative results were rarely replicated. But, they could occur when all of the following were true: [1] the paper was part of the “bushel” set in which all claims from the abstract were coded (~90% positive results), [2] some of the claims coded were negative results, and [3] a replication of that paper was able to include the negative results among the claims replicated. The two claims for which this occurred came from the same paper (ID: 7RR2).

- *Replication evidence for six claims were excluded because they fell far short of the planned sample size*. Sample size plans can fall short for several reasons. Some SCORE projects were severely affected by the COVID pandemic. The excluded papers (by ID) are: J40k, 9wya, jLr, 9wkl, J4W9, E4Am.

- We did not exclude all cases that fell short of power requirements. There were two occasions for conducting power estimates: [1] a priori estimates were conducted during research planning for peer review, [2] a statistical team conducted power analyses across all projects to assure consistency and address complex cases. These power estimates did not always agree.

    - For new data replication attempts, we included all cases in which both power estimates indicated that the requirements were met, and we included all cases in which the a priori power estimate indicated that the requirements were met. We also included all cases that were at least 80% of either the a priori target or the statistical team’s estimate in recognition that some variation around planned sample size occurs in the normal course of conducting research. We then excluded the 3 cases that did not meet any of these conditions as they were low powered regardless of estimation approach (paper IDs: Br0x, VvlD).

    - For secondary data replication attempts, we followed a similar methodology but without the lenience of achieving 80% power. We included all cases in which both power estimates indicated that the requirements were met, and we included all cases in which the a priori power estimate indicated that the requirements were met. We then excluded the 8 cases that did not meet any of these conditions as they were low powered regardless of estimation approach (paper IDs: AqDO, D2LY, AYQG, 9OK1, xGGO).

All together, 19 claims from 14 papers were excluded. For completeness, 5 of 17 of these claims (still excluding the two original negative results) showed replication success in terms of statistical significance and the same pattern.

Converting Statistical Outcomes to Common Effect Sizes

Comparing original and replication effect sizes has advantages over the binary assessments such as being a continuous representation of similarity and offering an emphasis on estimation rather than the presence or absence of an effect. In estimation contexts, the concept of power corresponds conceptually to the estimates' precision and is independent of the estimates' central tendencies. This increases confidence in comparing the average central tendencies for original and replication effect sizes across studies even if there is imprecise estimation for some individual studies. It also has disadvantages such as the lack of standardization of effect size metrics across statistical methods. There are some methods of converting effects in raw units to standardized units that can be applied across different models with few assumptions. But, there are some circumstances that require many assumptions about the comparability of the original and replication research designs and statistical models. There are other circumstances for which there is no possibility of establishing a meaningful, standardized metric.

Replicated papers and claims varied widely in research designs and statistical models. This resulted in a wide variety of native units for reporting the statistical outcomes. Part of the purpose of this project was to provide comparative results across the sample of replications. To achieve this purpose, it is helpful to have statistical outcomes reported on a common effect size metric. For some native units, conversions to common effect sizes is straightforward. For others, it is possible, but some assumptions must be made. And, for others, it is not possible to convert to effect size metrics in common with other replications. For reporting in the main text, we presented results aggregated to common effect size metrics to the extent possible. This involved some conversions with underlying assumptions that may or may not be met. Below, we summarize the conversions that were made, and summarize the results disaggregated by effect size metrics without conversions that require making meaningful assumptions.

In Table S7, we present disaggregated effect size calculations that require fewer assumptions but smaller samples for each metric. The qualitative interpretation of the overall results does not differ between aggregated and disaggregated approaches. Individual study outcomes are not the basis of conclusions in this paper, but any subsequent uses of these data for meta-analyses of individual outcomes should attend to variation based on effect size conversions.

Table S7. Selected ratios of effect sizes for original and replication studies in common units compared to Pearson's r partial correlation conversions

Effect size	Number of study pairs	Replication:original effect size ratio	Replication:original partial correlation ratio
Cohen's f-squared	28	0.84	0.39
Cohen's d	23	0.4	0.41
Correlation	8	0.32	0.24

Results

Replications completed by year in comparison with the sampling frame

In the main text, we reported changes in representativeness of the sample for replications attempted in comparison with the sample of claims *by discipline*. Table S8 provides the same

information *by year of publication*. The sampling process drew near equal numbers of papers eligible for replication for each year, and the distribution of papers with completed replications remained within 2.8% of the expectation of 10% per year except for just 6.1% for 2015. By number of unique claims replicated (last row), 2010 claims were most over-represented (from 9.0% [n = 352/3900] to 16.4% [n = 45/274]), and 2011 (from 9.2% [n = 357/3900] to 6.9% [n = 19/274]) and 2015 (from 10.4% [n = 404/3900] to 5.8% [n = 16/274]) claims were most underrepresented.

Table S8. Papers and claims selected, and replication attempts, by year of publication.

	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
						n (%)					
Claims selected											
Papers with claims	371 (9.5%)	352 (9.0%)	357 (9.2%)	388 (9.9%)	400 (10.3%)	391 (10.0%)	404 (10.4%)	409 (10.5%)	409 (10.5%)	419 (10.7%)	3900 (100%)
Papers eligible for replication	140 (9.3%)	114 (7.6%)	123 (8.2%)	152 (10.1%)	156 (10.4%)	148 (9.9%)	157 (10.5%)	166 (11.1%)	165 (11.0%)	179 (11.9%)	1500 (100%)
Papers with multiple claims	17 (8.5%)	18 (9.0%)	16 (8.0%)	13 (6.5%)	20 (10.0%)	26 (13.0%)	22 (11.0%)	25 (12.5%)	18 (9.0%)	25 (12.5%)	200 (100%)
Papers with single claim	123 (9.5%)	96 (7.4%)	107 (8.2%)	139 (10.7%)	136 (10.5%)	122 (9.4%)	135 (10.4%)	141 (10.8%)	147 (11.3%)	154 (11.8%)	1300 (100%)
Replications attempted											
Papers with replication started	25 (12.6%)	18 (9.1%)	20 (10.1%)	16 (8.1%)	21 (10.6%)	23 (11.6%)	16 (8.1%)	14 (7.1%)	22 (11.1%)	23 (11.6%)	198 (100%)
Papers with replication attempts completed	21 (12.8%)	16 (9.8%)	19 (11.6%)	12 (7.3%)	15 (9.1%)	19 (11.6%)	10 (6.1%)	13 (7.9%)	19 (11.6%)	20 (12.2%)	164 (100%)
Total replication attempts of claims	33 (11.1%)	49 (16.6%)	20 (6.8%)	39 (13.2%)	31 (10.5%)	28 (9.5%)	17 (5.7%)	22 (7.4%)	28 (9.5%)	29 (9.8%)	296 (100%)
Unique claims with replication attempts	32 (11.7%)	45 (16.4%)	19 (6.9%)	37 (13.5%)	26 (9.5%)	24 (8.8%)	16 (5.8%)	21 (7.7%)	26 (9.5%)	28 (10.2%)	274 (100%)

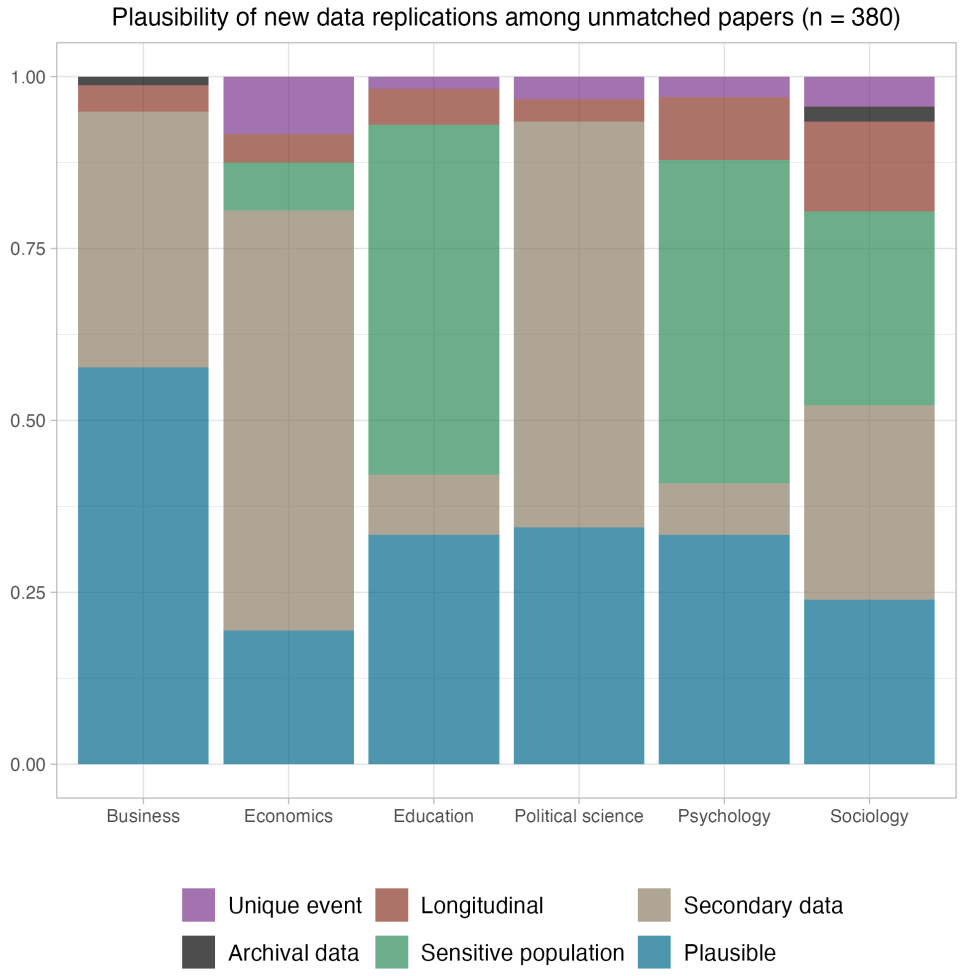
Papers from Phase 1 for which a replication was not attempted

The dataset of 600 papers created in Phase 1 remained representative of the sampling frame of quantitative empirical papers from 2009 to 2018 from 62 journals across the social-behavioral sciences. However, which of those papers ultimately had a replication attempt for one of its claims was contingent on available resources and expertise to conduct a good-faith replication. The main text summarizes shifts in representativeness from the sampling frame to the completed replication studies. Here, we provide further insight by providing some evidence for why papers did not get selected for replication. With infinite resources and time, a much higher proportion of papers would have been replicated, but the constraints of time, resources, and sourcing of data, experts, and instrumentation constrained the possibility of starting and completing replication studies.

Within the constrained set of 600 papers, we were able to evaluate reasons that replications were not attempted. 380 (63.3%) were not matched with a team to plan and conduct a replication study. After data collection was completed, we evaluated the papers and claims for potential barriers to replication feasibility as a new data replication (Figure S10) or as a secondary data replication (Figure S11). The Figures present that data by discipline as the

reasons vary substantially. For plausibility of conducting new data replications (Figure S10), 132 of the 380 papers (34.7%) were rated as plausible and the rest were deemed implausible because 132 (34.7%) were more likely conducted as secondary data replications, 78 (20.5%) were based on sensitive populations that are more difficult to obtain with time and resource constraints and challenging for the requirement to obtain secondary IRB approval through the Department of Defense (e.g., prison populations), 23 (6.1%) were longitudinal studies that exceeded the time window for data collection, and 13 (3.4%) were based on unique events to which the claim was directly relevant or for which similar events for conducting a replication would be challenging to identify. As could be anticipated, sensitive populations were more common reasons for education and psychology, and secondary data was a more common reason for economics, political science, and business.

Figure S10. Retrospective review of papers that were not matched to replication teams to conduct a new data replication by discipline

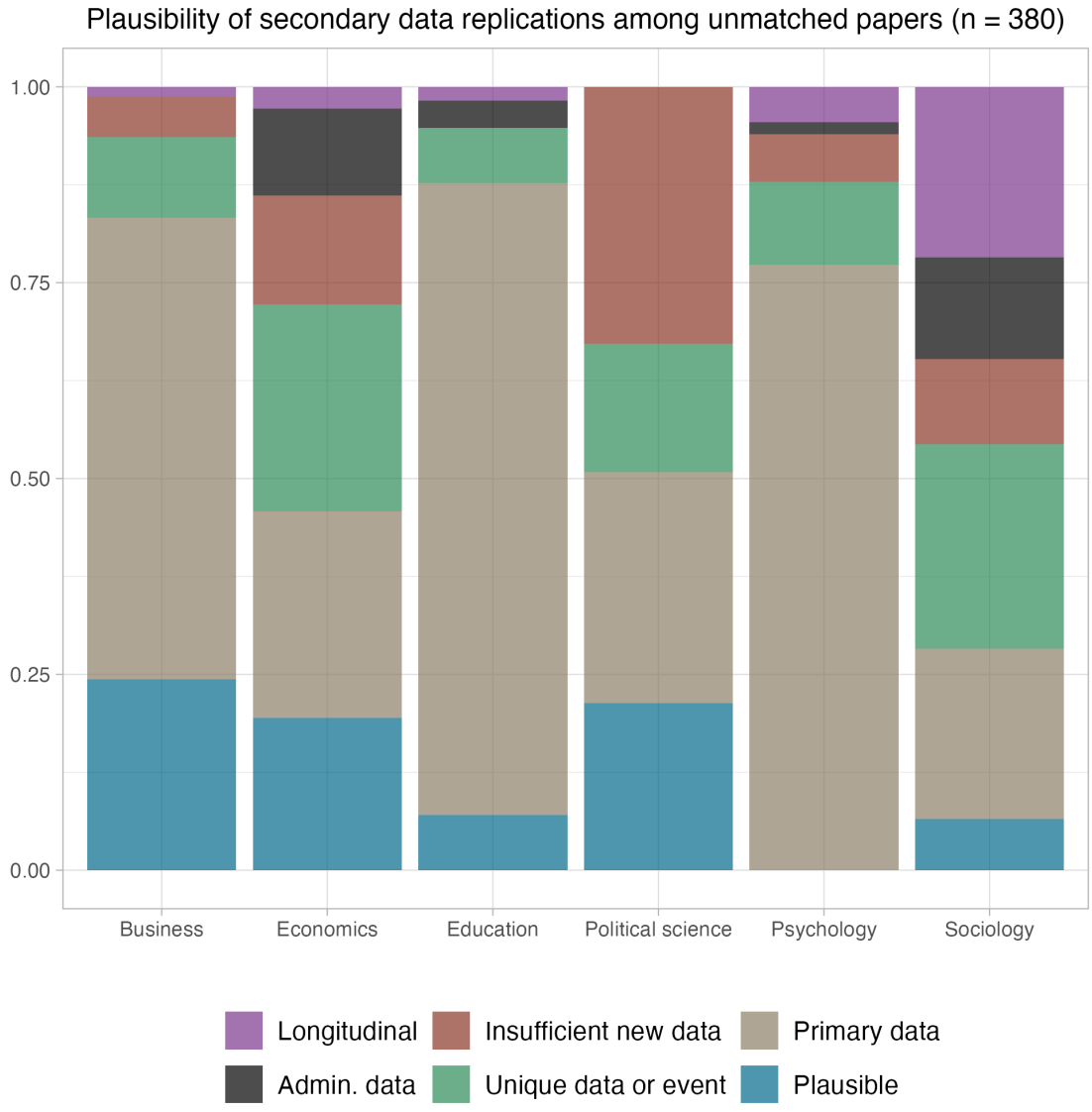


Caption: "Plausible" means that there were no clear barriers to conducting a replication other than capacity within the project. "Secondary data" means that these papers were more appropriate for a secondary data replication.

For plausibility of conducting secondary data replications of the same papers (Figure S11), 53 of the 380 papers (13.9%) were rated as plausible and the rest were deemed implausible because 190 (50.0%) were more likely conducted as new data replications, 60 (15.8%) were based on unique data or events for which it was unlikely to be feasible to find alternative data suitable to

conduct a good faith replication, 43 (11.3%) were unlikely to have sufficient data because, for example, the time period examined in the original study was longer than the time period that had since elapsed to conduct a fair replication of the original design, 17 (4.5%) were based on administrative data that is difficult to obtain, and 17 (4.5%) were based on longitudinal data for which it was unlikely there was sufficient independent data to conduct a replication. Figure S9 illustrates, for example, that insufficient new data was most common in political science, whereas unique data or events were common in economics, and all barriers were well represented in sociology.

Figure S11. Retrospective review of papers that were not matched to replication teams to conduct a secondary data replication by discipline



Caption: "Plausible" means that there were no clear barriers to conducting a replication other than capacity within the project. "Primary data" means that these papers were more appropriate for new data replications.

In total, from Phase 1, 181 of the 380 papers (47.6%) were plausible to have conducted a new or secondary data replication in the context of the program, if only there were more capacity

available. Many of the other 199 (52.4%) rated as implausible might have been possible to replicate if there were no constraints on time, resources, or expertise.

Comparing replication rates between replications selected during Phase 1 and Phase 2

Of the 164 papers with replications, 139 came from the sample defined during Phase 1 of the program and 25 came from the sample defined during Phase 2 of the program.

Phase 1 included the Evidence set of 600 papers randomly drawn from the dataset of 3000 papers. For Phase 2, an additional 900 papers were randomly drawn from the same original database of 27407 papers to create a complementary annotation set for rating by the human and machine teams. From these 900 papers, we conducted 25 replications, selected for feasibility given the time constraints of Phase 2 before the conclusion of the program (see Abatayo et al. [2025] for details on sampling). Table S9 highlights the distribution of replication papers from Phase 1 and Phase 2 across disciplines.

Table S9. Paper-level count of replications completed from Phase 1 and Phase 2 by discipline.

	Phase 1 papers	**Phase 2 papers**
Business	27	9
Economics and finance	22	2
Education	10	3
Political science	12	3
Psychology and health	54	4
Sociology and criminology	14	4
Total	139	25

Significance and effect size for replications completed from Phase 1 sample

Considering only replication attempts from the Phase 1 sample, weighting by paper, 68.8 of 139 papers replicated had statistically significant findings with the same pattern as the original finding 49.5% [95% CI 43.2 - 55.4%], 12.0 had statistically significant findings with an opposing pattern 8.6% [95% CI 5.1 - 11.8%], and 57.2 replications showed a null effect 41.2% [95% CI 35.1 - 46.6%]. Unweighted by claim, 139 of 249 replicated claims had statistically significant findings with the same pattern 50.7% [95% CI 44.8 - 56.6%], 20 had statistically significant findings with the opposite pattern 7.3% [95% CI 4.8 - 11.0%], and 89 showed a null effect 32.5% [95% CI 27.2 - 38.2%].

Table S10. Original and replication findings by Pearson's r effect size by papers and claims.

	Papers (weighted)		**Claims (unweighted)**	
	Original	Replication	Original	Replication
Number of *r* effect sizes	132		224	
Median [IQR] sample size	201 [124.0]	510 [336.0]	236 [2566.5]	556 [2352.8]
Median Pearson's r effect size (SD)	0.26 (0.21)	0.12 (0.18)	0.25 (0.21)	0.13 (0.18)

Note: Sample for this table is the studies for which a Pearson's r could be calculated for the original and replication outcomes. Papers are weighted combinations of claims accounting for multiple claims replicated in some papers. SD=standard deviation, IQR=interquartile range.

As summarized in Table S10, among the 132 papers for which a Pearson's *r* could be calculated for the original and replication outcomes, the median original effect size was 0.26 (SD = 0.21) and the median replication effect size was 0.12 (SD = 0.18), or 55.3% smaller by median. Among the 224 claims (unweighted) for which a Pearson's *r* could be calculated, the median original effect size was 0.25 (SD = 0.21) and the median replication effect size was 0.13 (SD = 0.18), or 45.9% smaller by median.

Significance and effect size for replications completed from Phase 2 sample

Considering only replication attempts from the Phase 2 sample, 12.0 of 25 papers replicated had statistically significant findings with the same pattern as the original finding 48.0% [95% CI 31.2 - 62.5%], 4.0 had statistically significant findings with an opposing pattern 16.0% [95% CI 5.9 - 26.7%], and 9.0 replications showed a null effect 36.0% [95% CI 21.4 - 50.0%]. Note that for this sample, there is one claim per paper, so claim and paper-level analyses are identical.

As summarized in Table S11, among the 25 papers for which a Pearson's r could be calculated for the original and replication outcomes, the median original effect size was 0.18 (SD = 0.17) and the median replication effect size was 0.09 (SD = 0.13), or 51.1% smaller by median.

Table S11. Original and replication findings by Pearson's r effect size by papers and claims.

	Papers	
	Original	Replication
Number of R effect sizes	25	
Median [IQR] sample size	271 [133.0]	1062 [698.0]
Median Pearson's r effect size (SD)	0.18 (0.17)	0.09 (0.13)

Note: Sample for this table is the studies for which a Pearson's r could be calculated for the original and replication outcomes. Papers are weighted combinations of claims accounting for multiple claims replicated in some papers. SD=standard deviation, IQR=interquartile range.

Replication success rates across binary assessments

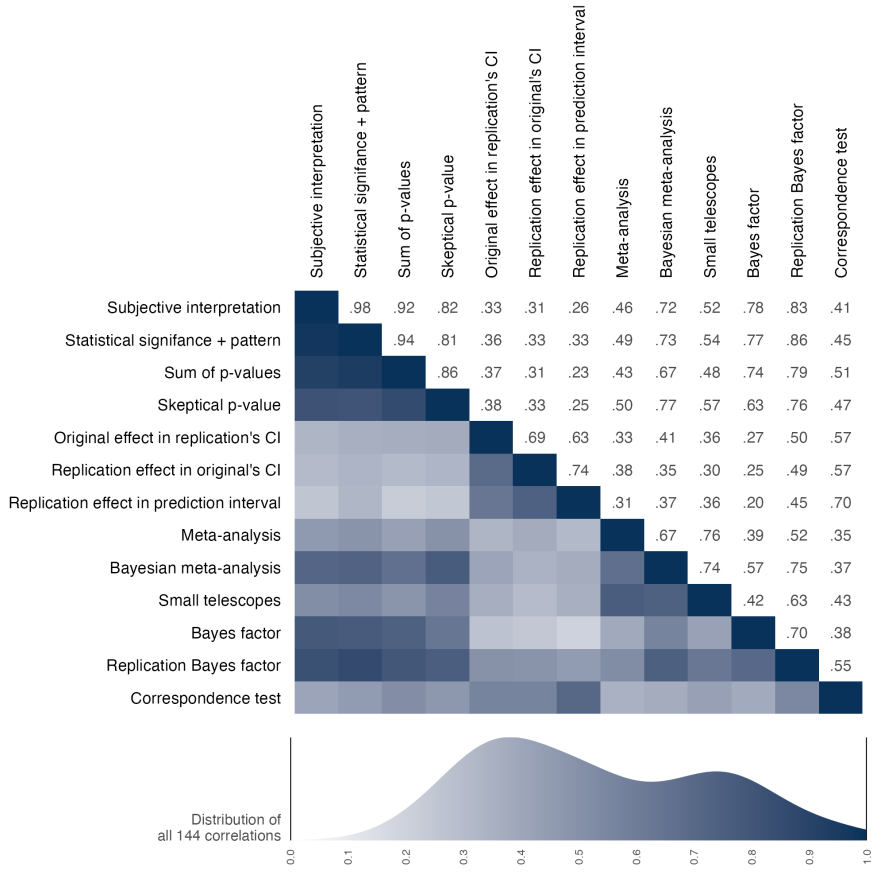
In the main text, we reported replication success rates across 13 binary assessments. Because of different assumptions, the binary assessments often were applicable to different subsets of the replication outcomes. As such, observed variation in replication success rates could be due both to different assumptions of each metric and the different subsets of findings to which the metrics were applied. Table S12 reports the replication success rates by claims for each of the binary assessments for all of the replication outcomes to which they could be applied (n's range 153 to 274) and to only the replication outcomes to which all of the binary assessments could be applied (n = 120).

Figure S12 shows the same correlation matrix as Figure 3 from the main text except the correlations are computed across claims rather than across papers. The pattern of correlations is generally consistent with the strength of correspondence observed between metrics across papers.

Table S12. Replication success rates for binary assessments by claims

	All claims possible		Complete cases (all 13 metrics)	
	Counts	Rate	Counts	Rate
Analyst interpretation	140 of 256	55%	59 of 120	49%
Bayes factor	97 of 240	40%	44 of 120	37%
Bayesian meta-analysis	148 of 225	66%	69 of 120	58%
Correspondence test	79 of 153	52%	60 of 120	50%
Meta-analysis	180 of 225	80%	90 of 120	75%
Orig. in rep. CI	86 of 249	35%	32 of 120	27%
Rep. in orig. CI	103 of 249	41%	36 of 120	30%
Rep. in prediction interval	125 of 225	56%	40 of 120	33%
Replication Bayes factor	122 of 225	54%	58 of 120	48%
Sig + pattern	151 of 274	55%	58 of 120	48%
Skeptical p-value	106 of 197	54%	63 of 120	52%
Small telescopes	185 of 248	75%	81 of 120	68%
Sum of p-values	95 of 206	46%	58 of 120	48%

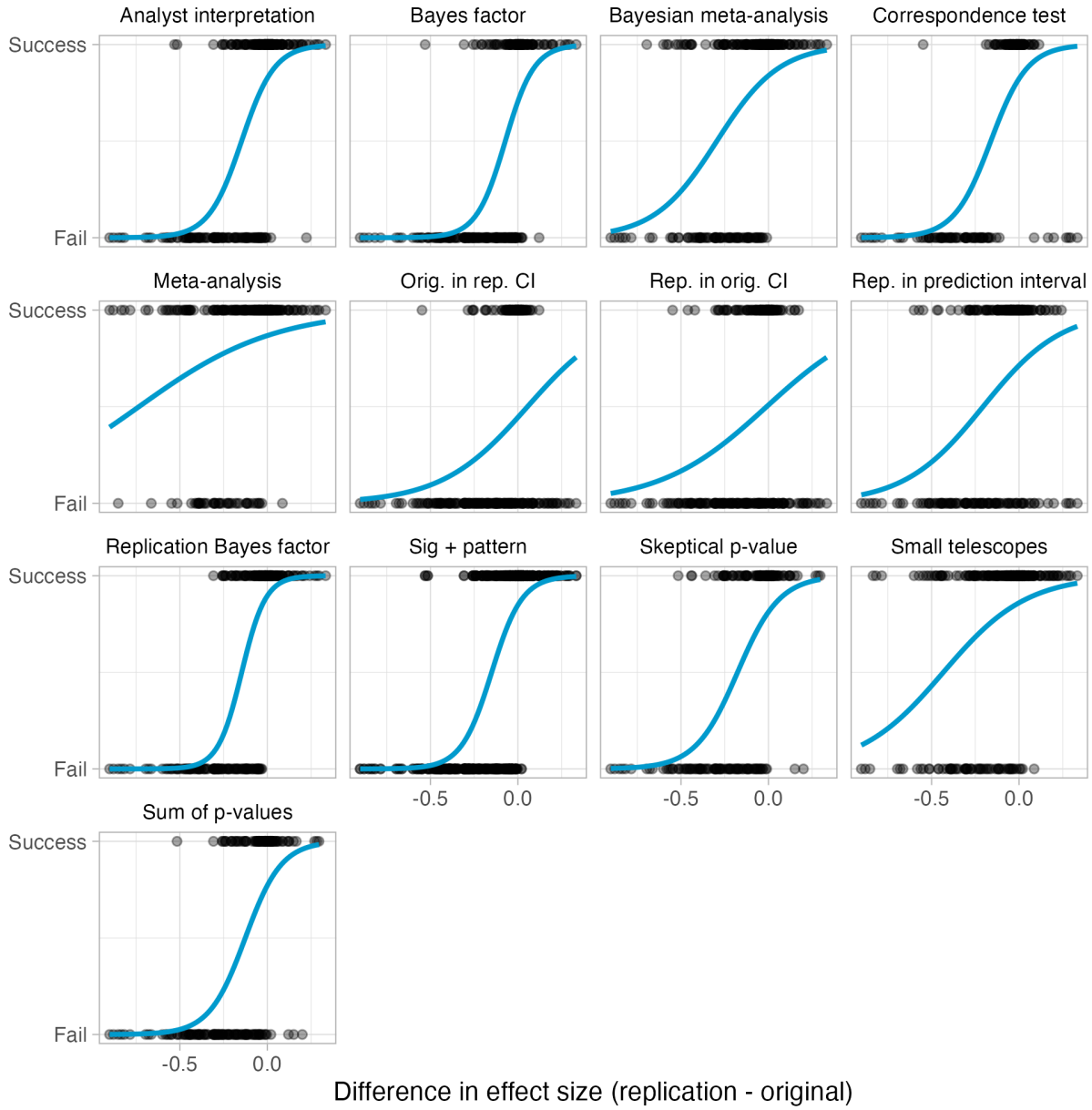
Figure S12: Correlation matrix among binary assessments of replication success across claims



Note: Correlation values are right of the diagonal, and correlation magnitude is visualized left of the diagonal with darker shading indicating stronger correlations. CI = confidence interval.

Figure S13 visualizes the pattern of success rates for relative effect sizes between replication and original studies for each of the 13 assessment methods with logistic curves. These patterns provide complementary insight for the correlations observed between assessments in Figure S12. For example, metrics with more failures when replication effect sizes were either higher or lower than original effect sizes show similar shapes.

Figure S13: Replication success or failure for 13 binary assessments by the effect size difference between the replication and original studies



Caption: Data points are differences in effect sizes for individual claims. Data points on top are successful replications, and data points on the bottom are failed replications, according to the metric.

Disciplinary differences in new versus secondary data replications

Table S13 shows the substantial differences by discipline in the proportion of replications that used new versus secondary data, median power to detect effect sizes 75% as large as the original study, and median sample sizes for original studies and replication attempts. Note that the relationships between power and sample size are not necessarily aligned as there is substantial variation in methodologies that can affect power estimation. The relationship between data source, discipline, and replication success deserves systematic attention in future research.

Table S13. Proportion of new versus secondary data replications and power to detect 75% of the original effect size by discipline

	Proportion of replication type		Median power to detect 75% of the original effect		Median sample size	
	New data	Secondary data	Assuming same scales	Assuming same designs	Original study	Replication attempt
Business	91.7%	8.3%	86.6%	47.7%	155.5	519
Economics and finance	45.8%	54.2%	67.7%	58.5%	189	276
Education	7.7%	92.3%	99.5%	76.9%	9189	8043
Political science	33.3%	66.7%	77.6%	58.1%	2236	2684
Psychology and health	81.0%	19.0%	67.3%	58.1%	107	437
Sociology and criminology	5.6%	94.4%	63.5%	42.6%	1860	1666

Replication outcomes by discipline

In the main text, we presented statistical significance and effect sizes by subdiscipline across papers. In Table S14, we reproduce the table with the same comparison across claims.

Table S14. Original and replication outcomes by statistical significance and pattern and by Pearson's *r* effect size for 6 subdisciplines across claims

	Replication attempt statistical significance and pattern		**Median Pearson's r effect size (SD)**	
	Counts	Percentage	Original	Replication
Business	17 of 36	47%	0.24 (0.12)	0.06 (0.16)
Economics and finance	18 of 38	47%	0.45 (0.2)	0.18 (0.26)
Education	20 of 28	71%	0.07 (0.22)	0.15 (0.14)
Political science	30 of 45	67%	0.16 (0.24)	0.15 (0.2)
Psychology and health	51 of 94	54%	0.26 (0.21)	0.09 (0.21)
Sociology and criminology	15 of 33	45%	0.13 (0.17)	0.02 (0.16)
Total	151 of 274	55%	--	--

Caption: SD=standard deviation

For the main text and the Table above, we aggregated the 62 journals into 6 disciplines. The original selection of journals was done considering representation from more subdisciplines that were aggregated to 6 for expository purposes. Social-behavioral subdisciplines have fuzzy boundaries, and journals do not necessarily abide by those boundaries in the content that they publish. Nevertheless, the selection of journals was based on considering nominations of journals that were representative of these subdisciplines to ensure diverse representation

across subfields. Here, we summarize the primary replication outcomes separating the 62 journals into smaller subdisciplines: 1 journals in Criminology, 7 journals in Economics, 7 journals in Education, 2 journals in Finance, 3 journals in Health, 4 journals in Management, 4 journals in Marketing, 3 journals in Organizational Behavior, 7 journals in Political Science, 10 journals in Psychology, 1 journals in Public Administration, and 6 journals in Sociology. Note, however, that subdiscipline boundaries are also fuzzy and several journals publish papers across those boundaries. These categorizations are imprecise given the actual content in the published papers.

Table S15 reproduces the findings reported above but separated out into 12 subdisciplines. An obvious caution is that the sample sizes for some of these subsets are quite small leading to highly imprecise results. There is not a strong basis for interpreting variation across subdisciplines as indicative of meaningful differences in replication success rates.

Table S15. Original and replication outcomes by statistical significance and pattern and by Pearson's *r* effect size for 12 subdisciplines across claims

	Statistical significance and pattern		**Median Pearson's r effect size (SD)**	
	Counts	Percentage	Original	Replication
Criminology	3 of 5	60%	0.13 (0.07)	0.07 (0.08)
Economics	17 of 34	50%	0.45 (0.18)	0.18 (0.27)
Education	20 of 28	71%	0.07 (0.22)	0.15 (0.14)
Finance	1 of 4	25%	0.61 (0.36)	0.14 (0.21)
Health	4 of 7	57%	0.08 (0.12)	0.09 (0.06)
Management	2 of 5	40%	0.16 (0.12)	0.01 (0.27)
Marketing	8 of 18	44%	0.24 (0.11)	0.03 (0.13)
Org. behavior	7 of 13	54%	0.24 (0.13)	0.12 (0.14)
Political science	29 of 44	66%	0.16 (0.24)	0.16 (0.2)
Psychology	47 of 87	54%	0.27 (0.21)	0.12 (0.21)
Public administration	1 of 1	100%	0.12 (NA)	0.11 (NA)
Sociology	12 of 28	43%	0.12 (0.18)	0.02 (0.17)

Caption: SD=standard deviation, Org=Organizational.

Original and Replication Effects by Year of Publication

A weighted chi-squared test on replication success across papers by publication year failed to reject the null hypothesis (p = 0.14). Differences in median estimates for effect sizes were largest for 2014 (original papers median R = 0.39; replications median R = 0.19) and smallest for 2016 (original papers median R = 0.14; replications median R = 0.11). Sample sizes for individual years are small, and there is no obvious trend by publication year for the relative magnitude of the difference between original and replication studies or their success rates in terms of statistical significance.

Across claims, the Pearson's *r* correlation between the proportion of statistical significant replications with the same pattern of the original finding with publication year was 0.064 [95% CI -0.084 - 0.214]. Differences in median estimates for effect sizes were largest for 2014 (original papers median R = 0.39; replications median R = 0.19) and smallest for 2016 (original papers median R = 0.14; replications median R = 0.11).

Sample sizes for individual years are small, and there is no obvious trend by publication year for the relative magnitude of the difference between original and replication studies or their success rates in terms of statistical significance.

Multiple replications of single claims

In most cases, a single replication was conducted on a single claim from a single paper. For some replication studies (total n = 28), we allowed multiple replications of a single claim. We offered this opportunity in case multiple teams were interested in the same claim and to involve as many replication teams as possible. Even so, we prioritized conducting replications of different claims, so there are not many of these cases. In the main text, the reported findings are based on the aggregated evidence across multiple replications of the same claim. Here, we provide insight to the individual replications comprising that aggregated evidence.

Same protocol

For 3 claims, multiple teams used the same protocol and conducted separate replication studies (n = 10 total studies). This mimics “Many Labs” replication projects (Ebersole et al., 2016, 2019; Klein et al., 2014, 2018, 2019). For these cases, a single preregistration protocol was created, peer reviewed, and approved. Then, multiple teams used that protocol to conduct their replication.

As shown in Table S16, 3 of the 10 individual data collections were directionally inconsistent (n = 1) or were directionally consistent but included 0 in their 95% confidence interval (n = 2). However, for all three of these findings, the cumulative evidence across the individual data collections provided a positive result consistent with the original finding.

Different protocol

For 12 claims (n = 24 total studies), multiple teams conducted replications of the same claim but used different protocols (Table S17). For these cases, each protocol was separately created, peer reviewed, and approved. This introduces greater heterogeneity between replications because the protocol, sample, and implementation context differed between the replications. These might show lower replicability rates than replication studies that make every effort to test the original claim as specified -- either by recreating the procedures as closely as possible, or by reproducing all of the key features believed to be necessary to observe evidence for the original claim (Nosek & Errington, 2020). Alternatively, if the claims are well-specified with the conditions necessary to observe them, then the variation in operationalization might be immaterial as long as that variation is irrelevant to the manifestation of the finding. Few claims in the social-behavioral sciences are so well-specified that variation in methods would not be associated with increasing heterogeneity in observed effect sizes.

Table S16. Outcomes of individual data collections of the same original claim (3 cases)

Paper	Journal	Original effect	Study ID	Re. sample	Replication effect
Yang et al. (2013)	Journal of Marketing Research	0.33 (0.14, 0.48)	887	169	0.15 (0, 0.29)
Yang et al. (2013)	Journal of Marketing Research	0.33 (0.14, 0.48)	369z	162	0.21 (0.05, 0.35)
Yang et al. (2013)	Journal of Marketing Research	0.33 (0.14, 0.48)	387	162	0.23 (0.08, 0.37)
Yang et al. (2013)	Journal of Marketing Research	0.33 (0.14, 0.48)	999g	168	-0.12 (-0.28, 0.04)
Yang et al. (2013)	**Journal of Marketing Research**	**0.33 (0.14, 0.48)**	**gz9m**	**--**	**0.12 (0.04, 0.2)**
King & Bryant (2017)	Journal of Organizational Behavior	0.27 (0.05, 0.45)	1642	183	0.36 (0.23, 0.47)
King & Bryant (2017)	Journal of Organizational Behavior	0.27 (0.05, 0.45)	om7	183	0.46 (0.34, 0.56)
King & Bryant (2017)	**Journal of Organizational Behavior**	**0.27 (0.05, 0.45)**	**40z0**	**--**	**0.42 (0.33, 0.51)**
Ku & Zaroff (2014)	Journal of Environmental Psychology	0.48 (0.3, 0.62)	4zy8	71	0.34 (0.12, 0.52)
Ku & Zaroff (2014)	Journal of Environmental Psychology	0.48 (0.3, 0.62)	15yz	81	0.34 (0.14, 0.51)
Ku & Zaroff (2014)	Journal of Environmental Psychology	0.48 (0.3, 0.62)	3z7z	152	0.05 (-0.12, 0.22)
Ku & Zaroff (2014)	Journal of Environmental Psychology	0.48 (0.3, 0.62)	8y1	150	0.14 (-0.02, 0.29)
Ku & Zaroff (2014)	**Journal of Environmental Psychology**	**0.48 (0.3, 0.62)**	**4930**	**--**	**0.17 (0.08, 0.26)**

Caption: Bolded rows denote the aggregated result. All effect sizes are partial correlations. The analysis for study 999g was revised prior to the meta-analysis to reflect the replication team's recommended data exclusions rather than their preregistered version. The preregistered version remains unchanged in the broader datasets.

Table S17. Outcomes of individual data collections of the same claim using different protocols

Paper	Journal	Original effect	Study ID	Rep. sample	Replication effect
Trémolière et al. (2012)	Cognition	0.28 (0.07, 0.45)	276	533	0 (-0.08, 0.09)
Trémolière et al. (2012)	*Cognition*	*0.28 (0.07, 0.45)*	*2y486*	*129*	*0.02 (-0.15, 0.19)*
Trémolière et al. (2012)	**Cognition**	**0.28 (0.07, 0.45)**	**6o4o5**	**--**	**0.02 (-0.05, 0.1)**
Rodriguez-Lara & Moreno-Garrido (2012)	Experimental Economics	0.65 (0.28, 0.81)	23m12	151	0.67 (0.57, 0.75)
Rodriguez-Lara & Moreno-Garrido (2012)	Experimental Economics	0.65 (0.28, 0.81)	895g	27	0.63 (0.28, 0.8)
Rodriguez-Lara & Moreno-Garrido (2012)	**Experimental Economics**	**0.65 (0.28, 0.81)**	**6o4z5**	**--**	**0.67 (0.56, 0.78)**
Alves et al. (2018)	Psychological Science	0.24 (0.1, 0.37)	24716	388	0.13 (0.03, 0.23)
Alves et al. (2018)	Psychological Science	0.24 (0.1, 0.37)	92g	361	0.25 (0.15, 0.36)
Alves et al. (2018)	**Psychological Science**	**0.24 (0.1, 0.37)**	**6917o**	**--**	**0.19 (0.12, 0.26)**
Bhattacharjee et al. (2017)	*Journal of Personality and Social Psychology*	*0.25 (0.14, 0.34)*	*2kgg2*	*656*	*-0.25 (-0.32, -0.18)*
Bhattacharjee et al. (2017)	Journal of Personality and Social Psychology	0.25 (0.14, 0.34)	7g66	364	0.41 (0.32, 0.49)
Bhattacharjee et al. (2017)	**Journal of Personality and Social Psychology**	**0.25 (0.14, 0.34)**	**65k41**	**--**	**-0.33 (-0.39, -0.28)**
Pastötter et al. (2013)	Cognition	0.29 (0.18, 0.38)	6738	705	0.24 (0.16, 0.31)
Pastötter et al. (2013)	*Cognition*	*0.29 (0.18, 0.38)*	*6m396*	*687*	*0.28 (0.21, 0.35)*
Pastötter et al. (2013)	**Cognition**	**0.29 (0.18, 0.38)**	**2kk4o**	**--**	**0.26 (0.21, 0.31)**
Griffiths & Tenenbaum (2011)	*Journal of Experimental Psychology: General*	*0.55 (0.28, 0.71)*	*288g2*	*477*	*0.55 (0.49, 0.61)*
Griffiths & Tenenbaum (2011)	Journal of Experimental Psychology: General	0.55 (0.28, 0.71)	89z7	30	0.61 (0.4, 0.74)
Griffiths & Tenenbaum (2011)	**Journal of Experimental Psychology: General**	**0.55 (0.28, 0.71)**	**6914o**	**--**	**0.54 (0.51, 0.58)**
Fox (2009)	Journal of Labor Economics	0.04 (0.03, 0.06)	10g2	202476	0 (0, 0.01)
Fox (2009)	Journal of Labor Economics	0.04 (0.03, 0.06)	944y	361381	0 (0, 0.01)
Fox (2009)	**Journal of Labor Economics**	**0.04 (0.03, 0.06)**	**2kk9o**	**--**	**0 (-0.01, 0)**
Armon et al. (2013)	European Journal of Personality	0.04 (0, 0.08)	g79m	28731	-0.02 (-0.03, -0.01)
Armon et al. (2013)	European Journal of Personality	0.04 (0, 0.08)	m7g7	7796	-0.01 (-0.03, 0.01)
Armon et al. (2013)	**European Journal of Personality**	**0.04 (0, 0.08)**	**y7g1**	**--**	**-0.02 (-0.03, -0.01)**
McCarter et al. (2010)	Organizational Behavior and Human Decision Processes	0.27 (0.02, 0.47)	5g9	545	0.02 (-0.01, 0.04)
McCarter et al. (2010)	Organizational Behavior and Human Decision Processes	0.27 (0.02, 0.47)	6ow46	320	-0.02 (-0.08, 0.04)
McCarter et al. (2010)	**Organizational Behavior and Human Decision Processes**	**0.27 (0.02, 0.47)**	**243yy**	**--**	**0.01 (-0.01, 0.03)**
Luttrell et al. (2016)	*Journal of Experimental Social Psychology*	*0.17 (0.02, 0.31)*	*24316*	*370*	*0.14 (0.03, 0.24)*

Luttrell et al. (2016)	Journal of Experimental Social Psychology	0.17 (0.02, 0.31)	9kzy	443	0.18 (0.09, 0.27)
Luttrell et al. (2016)	**Journal of Experimental Social Psychology**	**0.17 (0.02, 0.31)**	**23my4**	**--**	**0.16 (0.09, 0.23)**
Boehm et al. (2015)	Psychological Science	-0.05 (-0.08, -0.02)	651m6	4971	-0.02 (-0.05, 0.01)
Boehm et al. (2015)	Psychological Science	-0.05 (-0.08, -0.02)	z516	7981	0.12 (0.06, 0.18)
Boehm et al. (2015)	**Psychological Science**	**-0.05 (-0.08, -0.02)**	**21zyw**	**--**	**0 (-0.02, 0.03)**
Ohtsubo et al. (2014)	*Evolution and Human Behavior*	*0.75 (0.54, 0.85)*	*128g6*	*41*	*0.17 (-0.14, 0.44)*
Ohtsubo et al. (2014)	Evolution and Human Behavior	0.75 (0.54, 0.85)	9977	38	-0.07 (-0.37, 0.25)
Ohtsubo et al. (2014)	**Evolution and Human Behavior**	**0.75 (0.54, 0.85)**	**2wgy9**	**--**	**0.06 (-0.16, 0.27)**

Caption: Bolded rows denote the aggregated result. Italicized rows denote generalizability studies. All effect sizes are partial correlations. The focal analysis for study 24316 was redesigned prior to the meta-analysis to bring it more in line with the original study. Consequently, the individual outcome appears as a successful replication here but as an unsuccessful replication in the reported data.

Hybrid replications

We defined hybrid replications as those that were not based on completely independent data. A common case for this is a longitudinal analysis in which additional waves of data are available since the original study. In 27 cases, replications used some of the original data and added new data. For 11 of those, a “pure” replication is reported with only independent data in the main text, and for 16 there was insufficient data to include an independent replication. Hybrids were not reported in the main text because they are not independent tests of the original claim.

Here, we summarize the hybrid versions that included original and new data together. Table S18 reports the individual effects for 27 cases that had hybrid data. For the 11 cases that had sufficient independent data to include a replication attempt in the main paper, the effects for the hybrid data (labeled “Hybrid” in the analysis column) are immediately followed by the effects with the independent data (labeled “Replication” in the analysis column).

Among the 16 cases that had only hybrid data, 8 succeeded and 8 failed to observe statistically significant evidence with the same pattern as the original study (50% success rate). Among the 11 cases for which there was sufficient data to report an independent replication, 5 succeeded with both hybrid and independent data (45%), 3 failed with both hybrid and independent data (27%), and 3 succeeded with hybrid data and failed with independent data (27%). Overall, 16 of 27 hybrid datasets successfully replicated the original study (59%), and the 5 of 11 independent datasets from this sample did so (45%).

Table S18. Effects for hybrid replication attempts that included original and independent data

Paper	Journal	Claim ID	Original effect	Study ID	Analysis	Rep. sample	Replication effect	SCORE outcome
0PZl	Comparative Political Studies	0PZl_single-trace	0.89 (0.84, 0.92)	g241	Hybrid	83	0.95 (0.93, 0.96)	Success
2GKO	World Development	2GKO_3pnxyo	0.19 (0.05, 0.32)	6m34m	Hybrid	452	-0.05 (-0.15, 0.04)	Failed
2GKO	World Development	2GKO_jr67qq	0.24 (0.11, 0.36)	6m34m	Hybrid	452	-0.18 (-0.26, -0.09)	Failed
2GKO	World Development	2GKO_pjp7lx	0.19 (0.05, 0.32)	6m34m	Hybrid	452	-0.06 (-0.15, 0.03)	Failed
2GKO	World Development	2GKO_single-trace	0.19 (0.05, 0.32)	4142	Hybrid	468	-0.05 (-0.14, 0.04)	Failed
2GKO	World Development	2GKO_single-trace	0.19 (0.05, 0.32)	4142	Replication	256	-0.17 (-0.28, -0.05)	Failed
2GKO	World Development	2GKO_wl74j8	0.19 (0.05, 0.32)	6m34m	Hybrid	452	0.07 (-0.02, 0.16)	Failed
2GKO	World Development	2GKO_x76512	0.2 (0.07, 0.33)	6m34m	Hybrid	452	-0.24 (-0.32, -0.15)	Failed
3aPw	Demography	3aPw_single-trace	--	05g8	Hybrid	420530	--	Success
3aPw	Demography	3aPw_single-trace	--	05g8	Replication	161855	--	Success
4q0L	Social Forces	4q0L_single-trace	0.22 (0.1, 0.32)	3z5z	Hybrid	500	0.18 (0.09, 0.27)	Success
4q0L	Social Forces	4q0L_single-trace	0.22 (0.1, 0.32)	3z5z	Replication	150	0.21 (0, 0.39)	Success
5Kgq	Demography	5Kgq_single-trace	-0.43 (-0.56, -0.28)	k6y7	Hybrid	165	-0.41 (-0.52, -0.28)	Success
5Kgq	Demography	5Kgq_single-trace	-0.43 (-0.56, -0.28)	k6y7	Replication	159	0.03 (-0.13, 0.18)	Failed
7X54	American Sociological Review	7X54_single-trace	0.33 (0.29, 0.37)	93k7	Hybrid	3347	0.35 (0.32, 0.38)	Success
7X54	American Sociological Review	7X54_single-trace	0.33 (0.29, 0.37)	93k7	Replication	1438	0.4 (0.36, 0.44)	Success
9wya	American Political Science Review	9wya_single-trace	0.09 (0.02, 0.15)	9k2y	Hybrid	1324	0 (-0.05, 0.06)	Failed
BKxK	Journal of Financial Economics	BKxK_single-trace	0.36 (0, 0.61)	4zz0	Hybrid	30	0.48 (0.14, 0.68)	Success
BKxK	Journal of Financial Economics	BKxK_single-trace	0.36 (0, 0.61)	4zz0	Replication	30	0.23 (-0.13, 0.52)	Failed
BebG	Psychological Science	BebG_single-trace	0.05 (0, 0.11)	42k8	Hybrid	2265	0.07 (0.03, 0.12)	Success
E5qr	World Development	E5qr_single-trace	0.25 (0, 0.46)	95y	Hybrid	114	-0.07 (-0.25, 0.12)	Failed

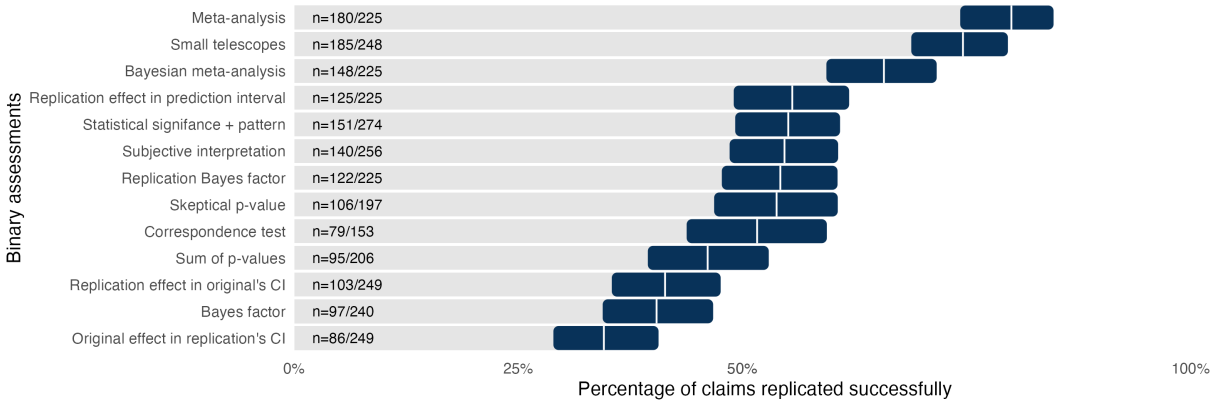
EJpm	American Journal of Sociology	EJpm_single-trace	0.35 (0.27, 0.41)	yk20	Hybrid	713	0.19 (0.12, 0.25)	Success
Ryq7	British Journal of Political Science	Ryq7_single-trace	0.54 (0.14, 0.75)	95my	Hybrid	29	0.42 (0.07, 0.65)	Success
VOwm	Management Science	VOwm_single-trace	0.1 (0.04, 0.16)	ykk0	Hybrid	32667	--	Failed
VOwm	Management Science	VOwm_single-trace	0.1 (0.04, 0.16)	ykk0	Replication	12344	-0.02 (-0.07, 0.03)	Failed
YRvg	Journal of Financial Economics	YRvg_single-trace	0.08 (-0.01, 0.17)	3g4k	Hybrid	641	0.12 (0.04, 0.19)	Success
kXp8	Organization Science	kXp8_single-trace	0.24 (0.11, 0.36)	kzyz	Hybrid	205	0.16 (0.04, 0.27)	Success
kXp8	Organization Science	kXp8_single-trace	0.24 (0.11, 0.36)	kzyz	Replication	174	0.15 (0.03, 0.27)	Success
l22v	Comparative Political Studies	l22v_single-trace	0.16 (0.03, 0.28)	68	Hybrid	99	0.13 (0.04, 0.22)	Success
q8xv	Journal of Political Economy	q8xv_single-trace	-0.54 (-0.66, -0.37)	mkk9	Hybrid	786	-0.52 (-0.63, -0.37)	Success
q8xv	Journal of Political Economy	q8xv_single-trace	-0.54 (-0.66, -0.37)	mkk9	Replication	391	-0.54 (-0.65, -0.39)	Success
qgWj	British Journal of Political Science	qgWj_single-trace	0.48 (0.28, 0.62)	2yg	Hybrid	95470	0.09 (0.08, 0.1)	Success
qgWj	British Journal of Political Science	qgWj_single-trace	0.48 (0.28, 0.62)	2yg	Replication	45119	0.2 (-0.13, 0.47)	Failed
qkNz	Journal of Labor Economics	qkNz_single-trace	0.43 (0.17, 0.61)	0y68	Hybrid	64	0.26 (0.01, 0.46)	Success
vmxO	World Development	vmxO_single-trace	0.12 (0.07, 0.17)	y4y0	Hybrid	2051	0.04 (0, 0.09)	Failed
z0v1	Social Forces	z0v1_single-trace	0.01 (0, 0.03)	0056	Hybrid	16832	0.03 (0.01, 0.04)	Success
zmYY	Criminology	zmYY_single-trace	0.13 (0.07, 0.19)	g5m	Hybrid	1409	0.03 (0, 0.06)	Failed
zmYY	Criminology	zmYY_single-trace	0.13 (0.07, 0.19)	g5m	Replication	1384	0 (-0.03, 0.03)	Failed

Caption: Studies for which there was sufficient data to report an independent replication in the main text are reported in two rows, the first row reporting the hybrid effects and the second row reporting the independent replication. Empty cells for 'Original effect' and 'Replication effect' mean the effect could not be reliably converted to a partial correlation. The SCORE outcome labels the effect "success" if it was statistically significant ($p < .05$) with the same pattern as the original study, and "failure" if it was not.

Success Rates on Binary Assessments Across Claims

In the main text, we reported the replication success rates across 13 binary assessments weighting claims by paper. In Figure S14, we report the same findings unweighted by claims.

Figure S14. Replication success rates across 13 binary assessments for claims



Caption: The vertical white line for each row is the estimate, and the 95% confidence interval around the estimate is represented by the dark bar. CI = confidence interval.

Topics and methodologies represented in the replicated papers

A reviewer expressed interest in more information about the topics and methodologies represented in the replicated papers. Full project descriptions are available on OSF, but we also created summaries of the papers by using large language models (LLMs) and topic modeling to summarize information based on original paper titles and abstracts.

We did not verify the accuracy of these summaries with human coding. They are provided for general impressions of the content of the replicated papers. Any inferential research with paper contents should be conducted systematically with accuracy validation.

Table S19 presents 13 themes extracted from the papers following the prompt to GPT-4.1: “What is the main research question?” We then generated text embeddings of each paper’s research question using Qwen3-Embeeding-8B, and performed topic modeling on these embeddings using the Python library BERTopic, which resulted in 13 themes with at least 6 papers fitting into each of these themes. The theme label/question was generated by GPT-4.1, which was provided with four representative papers’ research questions and was prompted to generate an overarching research question.

Table S19. Themes extracted from titles and abstracts of replicated papers.

Theme number	Number of papers	Theme
1	25	How do perceptions of profit, self-interest, and contextual framing shape individuals' moral reasoning, fairness judgments, and economic decision-making in social and experimental settings?
2	15	How do language minority status, bilingualism, socioeconomic factors, and neighborhood disadvantage interact to influence English reading achievement and behavioral development trajectories among elementary school children in the U.S.?
3	14	How do learned experience, linguistic factors, and individual cognitive or affective traits interact to shape perceptual representations, attentional mechanisms, and the learnability of complex hierarchical structures in humans?
4	12	How do electoral systems, local conflict dynamics, and candidate strategies shape patterns of electoral manipulation, party system structure, and the alignment between voter preferences and elected representatives in diverse political environments?
5	12	How do consumer motivations, self-concept, and contextual or material factors interact with marketing practices and product cues to shape perceptions, decision fatigue, preferences, and the desirability of products, including counterfeit goods?
6	12	How do health behaviors, social factors, and psychosocial processes interact with individual characteristics—such as sexual orientation, substance use, and cultural context—to shape mental and physical health outcomes across diverse populations?
7	11	How do family structure, socio-economic context, and cultural or demographic factors shape union formation, marital transitions, and household dynamics, and what are the implications for individual health and gendered household labor across societies and over time?
8	9	How do various market frictions—including margin requirements, funding constraints, short-sale impediments, and risks related to cash flow and volatility—affect cross-sectional differences and predictability in asset prices and stock returns, particularly during periods of financial stress or heightened investor sentiment?
9	8	How do individual differences in motivation, executive function, and self-efficacy interact with contextual factors and psychological states (such as mortality salience and mental contrasting) to shape cognitive effort, creative behavior, and moral decision-making?
10	7	How do racial and intersecting social identities shape detection, perception, evaluation, and implicit attitudes within psychological and societal contexts, particularly regarding underidentification, prejudice, stereotype influence, and intergroup favoritism?
11	7	How do individual differences in traits such as optimism and narcissism, as well as perceptions of social support and ideal partner matching, interact to influence psychological well-being and romantic relationship outcomes?
12	6	How do individual differences (such as family structure, age, psychological resilience) and non-work factors (e.g., after-work activities, recovery experiences, work desires) shape employees' work absorption, affective states, and preferences for job features across different professional contexts?
13	6	How do immigration, economic integration, and external financial influences shape public attitudes toward redistribution, welfare policies, and political trust in Europe and its neighboring regions, considering mediating factors like perceived threat, political mobilization, and historical context?

Note. 29 papers could not be assigned to any of the themes.

Figure S15 highlights the percentage of papers that were identified as using each of the listed statistical techniques or analytic approaches, which were automatically extracted by two LLMs (GPT-4.1 and Kimi K2). These results follow the prompt: "What statistical techniques or analytic approaches are used?" Note that the LLMs inferred the use of these techniques from information available in the abstracts.

Figure S15. Percentage of replicated papers that were automatically identified as using each method or technique. Two LLMs (GPT-4.1 and Kimi K2) identified the range of methods or techniques used across all abstracts (prompt: “What statistical techniques or analytic approaches are used?”). They then coded each abstract for the presence (1) or absence (0) of each—a method/technique is considered present if at least one of the models identified it as being present.

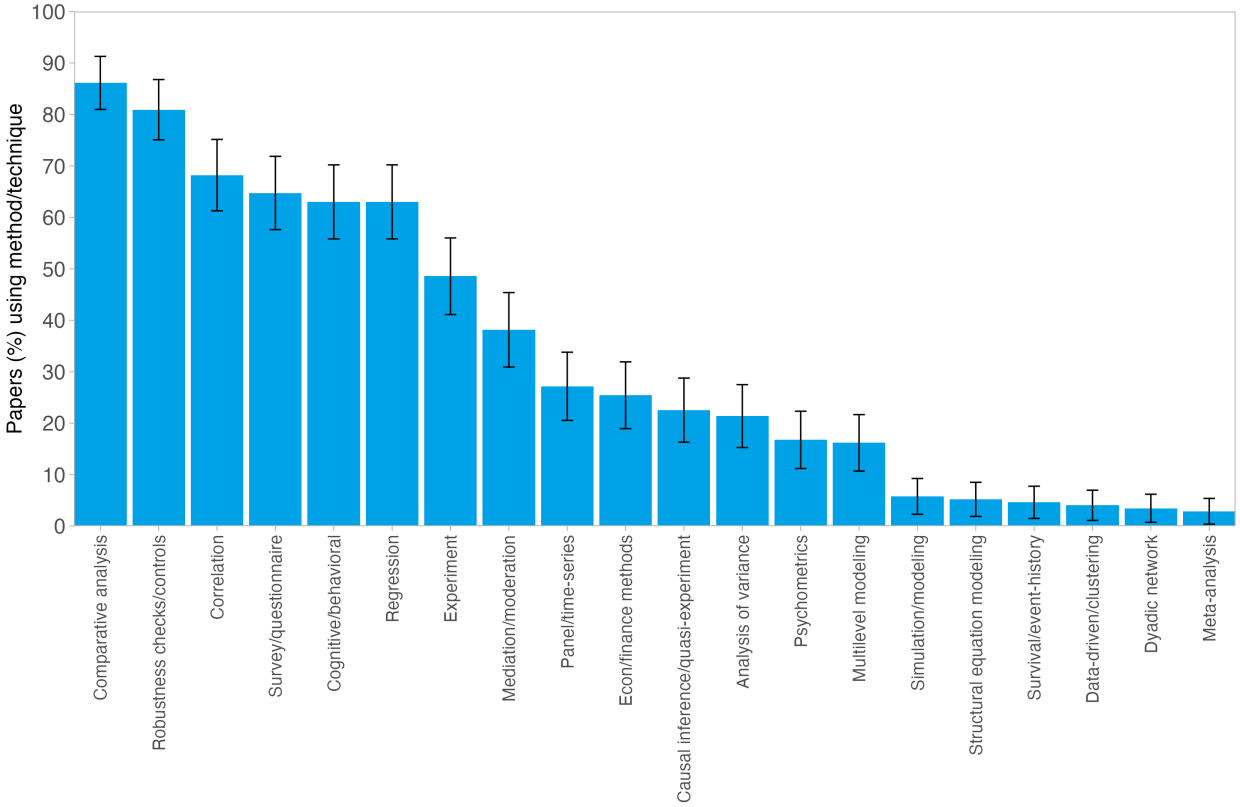
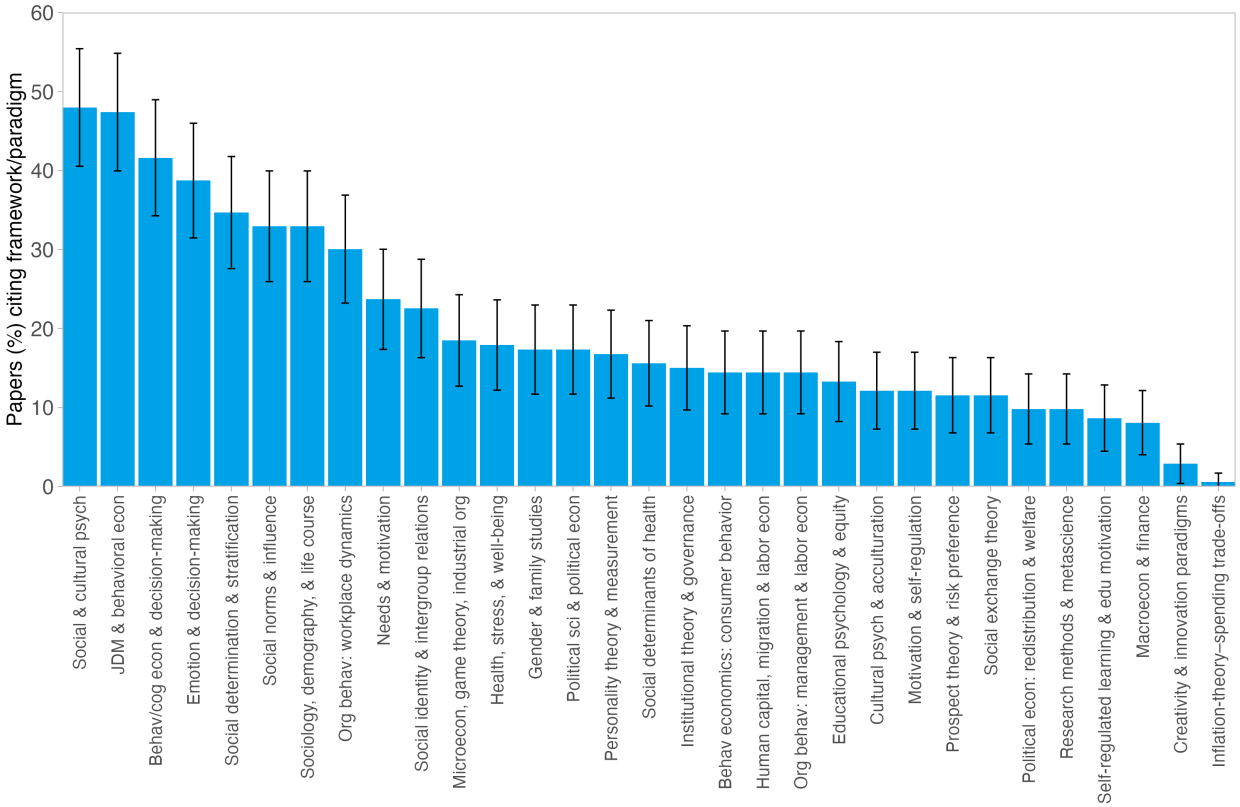


Figure S16 highlights the percentage of papers that were identified as using each of the listed theoretical frameworks, which were automatically extracted by two LLMs (GPT-4.1 and Kimi K2). These results follow the prompt: “What are the main theoretical frameworks/paradigms being cited?” Note that the LLMs inferred the use of these techniques from information available in the abstracts.

Figure S16. Percentage of replicated papers that were automatically identified as citing each theoretical framework or paradigm. Two LLMs (GPT-4.1 and Kimi K2) identified the range of frameworks or paradigms used across all abstracts (prompt: “What are the main theoretical frameworks/paradigms being cited?”). They then coded each abstract for the presence (1) or absence (0) of each—a framework/paradigm is considered present if at least one of the models identified it as being present.



References

Abatayo, A. L., Achakulvisut, T., Acuna, D., Aczel, B., Balaji, L., Bandrowski, A. E., & Benjamin, D. M. (2025). Empirical, Human, and Machine Assessments of Research Credibility in the Social and Behavioral Sciences.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour, 2(9)*, 637–644. https://doi.org/10.1038/s41562-018-0399-z

Chambers, C. (2019). What’s next for Registered Reports? *Nature, 573(7773)*, 187–189. https://doi.org/10.1038/d41586-019-02674-6

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of Registered Reports. *Nature Human Behaviour, 6(1)*, Article 1. https://doi.org/10.1038/s41562-021-01193-7

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., Ford, C., Volcic, R., De Rosario, H., & De Rosario, M. H. (2018). Package 'pwr.' R Package Version, 1(2). http://14.63.219.55/web/packages/pwr/pwr.pdf

Ebersole CR, Atherton OE, Belanger AL, Skulborstad HM, Allen JM, et al. 2016a. Many Labs 3: evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology, 67*:68–82.

Ebersole CR, Mathur MB, Baranski E, Bart-Plange D-J, Buttrick NR, et al. 2020. Many Labs 5: testing pre-data-collection peer review as an intervention to increase replicability. *Advances in Methods and Practices in Psychological Science, 3(3*):309–31.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3(4)*, 486–504.

Klein RA, Cook CL, Ebersole CR, Vitiello C, Nosek BA, et al. (2024). Many Labs 4: failure to replicate mortality salience effect with and without original author involvement. *Collabra*. https://doi.org/10/ghwq2w

Klein RA, Ratliff KA, Vianello M, Adams RB, Bahník Š, et al. (2014). Investigating variation in replicability: a "many labs" replication project. *Social Psychology, 45(3)*:142–52.

Klein RA, Vianello M, Hasselman F, Adams BG, Adams RB, et al. (2018). Many Labs 2: investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science, 1(4)*:443–90.

Laird, N. M., & Mosteller, F. (1990). Some statistical methods for combining experimental results. *International Journal of Technology Assessment in Health Care, 6(1)*, 5–30.

Miske, O., Abatayo, A. L., Daley, M., Dirzo, M., Fox, N., Haber, N., Hahn, K., Mawhinney, B., Silverstein, P., Stankov, T., & Tyner, A. H. (2025). Investigating the reproducibility of the social and behavioral sciences.

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLOS Biology, 18(3)*, e3000691. https://doi.org/10.1371/journal.pbio.3000691