Estimating the Reproducibility of Psychological Science

Abstract

Reproducibility is a defining feature of science, but the extent to which it characterizes science is unknown. Following a structured protocol, we conducted 100 replications of studies published in three psychology journals. Replication effects (M = .198, SD = .255) were half the magnitude of original effects (M = .396, SD = .193) representing a substantial decline effect. Ninety-seven percent of original studies had significant results (p < .05). Thirty-six percent of replications had significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 38% of effects were subjectively rated to have replicated the original result; and, if no bias in original results is assumed, 70% of the meta-analytic combinations were significant effects. Correlational tests suggest that replication success is better predicted by the strength of original evidence than by characteristics of the original and replication teams. In sum, a large portion of replications did not reproduce evidence supporting the original results despite using high-powered designs and original materials when available. The open dataset provides a basis for hypothesis generation on the causes of irreproducibility.

Abstract word count = 174 words
Keywords = Reproducibility, psychology, replication, meta-analysis, decline effect
Authors' Note: Authors are listed alphabetically. This project was supported by the Center for Open Science and the Laura and John Arnold Foundation. The authors declare no financial conflict of interest with the reported research.

Group Author: Open Science Collaboration[1]

[1] Authors listed alphabetically: Alexander A. Aarts, Nuenen, the Netherlands; Christopher J. Anderson, Southern New Hampshire University; Joanna E. Anderson, Defence Research and Development Canada; Peter R. Attridge, Mercer School of Medicine, Georgia Gwinnett College; Angela Attwood, University of Bristol; Jordan Axt, University of Virginia; Molly Babel, University of British Columbia; Štěpán Bahník, University of Würzburg; Erica Baranski, University of California, Riverside; Michael Barnett-Cowan, University of Waterloo; Elizabeth Bartmess, University of California, San Francisco; Jennifer Beer, University of Texas at Austin; Raoul Bell, Heinrich Heine University Düsseldorf; Heather Bentley, Georgia Gwinnett College; Leah Beyan, Georgia Gwinnett College; Grace Binion, University of Oregon, Georgia Gwinnett College; Denny Borsboom, University of Amsterdam; Annick Bosch, Radboud University Nijmegen; Frank A. Bosco, Virginia Commonwealth University; Sara D. Bowman, Center for Open Science; Mark J. Brandt, Tilburg University; Erin Brasell, Center for Open Science; Hilmar Brohmer, Tilburg University; Benjamin T.

Brown, Georgia Gwinnett College; Kristina Brown, Georgia Gwinnett College; Jovita Brüning, Humboldt University of Berlin, Charité - Universitätsmedizin Berlin; Ann Calhoun-Sauls, Belmont Abbey College; Shannon P. Callahan, University of California, Davis; Elizabeth Chagnon, University of Maryland; Jesse Chandler, University of Michigan, Mathematica Policy Research; Christopher R. Chartier, Ashland University; Felix Cheung, Michigan State University, University of Hong Kong; Cody D. Christopherson, Southern Oregon University; Linda Cillessen, Radboud University Nijmegen; Russ Clay, College of Staten Island, City University of New York; Hayley Cleary, Virginia Commonwealth University; Mark D. Cloud, Lock Haven University; Michael Cohn, University of California, San Francisco; Johanna Cohoon, Center for Open Science; Simon Columbus, University of Amsterdam; Andreas Cordes, University of Göttingen; Giulio Costantini, University of Milan-Bicocca; Leslie D. Cramblet Alvarez, Adams State University; Ed Cremata, University of Southern California; Jan Crusius, University of Cologne; Jamie DeCoster, University of Virginia; Michelle A. DeGaetano, Georgia Gwinnett College; Nicolás Della Penna, Australian National University; Bobby den Bezemer, University of Amsterdam; Marie K. Deserno, University of Amsterdam; Olivia Devitt, Georgia Gwinnett College; Laura Dewitte, University of Leuven; David G. Dobolyi, University of Virginia; Geneva T. Dodson, University of Virginia; M. Brent Donnellan, Texas A & M; Ryan Donohue, Elmhurst College; Rebecca Dore, University of Virginia; Angela Dorrough, University of Siegen, Max Planck Institute for Research on Collective Goods; Anna Dreber, Stockholm School of Economics; Michelle Dugas, University of Maryland; Elizabeth W. Dunn, University of British Columbia; Kayleigh Easey, Bristol University; Sylvia Eboigbe, Georgia Gwinnett College; Casey Eggleston, University of Virginia; Jo Embley, University of Kent; Sacha Epskamp, University of Amsterdam; Timothy M. Errington, Center for Open Science; Vivien Estel, University of Erfurt; Frank J. Farach, University of Washington, Prometheus Research; Jenelle Feather, Massachusetts Institute of Technology; Anna Fedor, Parmenides Center for the Study of Thinking; Belén Fernández-Castilla, Universidad Complutense de Madrid; Susann Fiedler, Max Planck Institute for Research on Collective Goods; James G. Field, Virginia Commonwealth University; Stanka A. Fitneva, Queen's University; Taru Flagan, University of Texas, Austin; Amanda L. Forest, University of Pittsburgh; Eskil Forsell, Stockholm School of Economics; Joshua D. Foster, University of South Alabama; Michael C. Frank, Stanford University; Rebecca S. Frazier, University of Virginia; Heather Fuchs, University of Cologne; Philip Gable, University of Alabama; Jeff Galak, Carnegie Mellon University; Elisa Maria Galliani, University of Padua; Anup Gampa, University of Virginia; Sara Garcia, Universidad Nacional De Asunción; Douglas Gazarian, Bard College; Elizabeth Gilbert, University of Virginia; Roger Giner-Sorolla, University of Kent; Andreas Glöckner, University of Göttingen, Max Planck Institute for Research on Collective Goods; Lars Goellner, University of Siegen; Jin X. Goh, Northeastern University; Rebecca Goldberg, Mississippi State University; Patrick T. Goodbourn, University of Sydney; Shauna Gordon-McKeon, Hampshire College; Bryan Gorges, Center for Open Science; Jessie Gorges, Center for Open Science; Justin Goss, Colorado State University-Pueblo; Jesse Graham, University of Southern California; James A. Grange, Keele University; Jeremy Gray, Michigan State University; Chris Hartgerink, Tilburg University; Joshua Hartshorne, Massachusetts Institute of Technology; Fred Hasselman, Radboud University Nijmegen, School of Pedagogy and Educational Science & Behavioural Science Institute: Learning and Plasticity; Timothy Hayes, University of Southern California; Emma Heikensten, Stockholm School of Economics; Felix Henninger, University of Koblenz-Landau, Max Planck Institute for Research on Collective Goods; John Hodsoll, NIHR Biomedical Research Centre for Mental Health at the South London, Maudsley NHS Foundation Trust, King's College London; Taylor Holubar, Stanford University; Gea Hoogendoorn, Tillburg University; Denise J. Humphries, Georgia Gwinnett College; Cathy O.-Y. Hung, University of Hong Kong; Nathali Immelman, University of Winchester; Vanessa C. Irsik, University of Nevada, Las Vegas; Georg Jahn, University of Lübeck; Frank Jäkel, University of Osnabrück; Marc Jekel, University of Göttingen; Magnus Johannesson, Stockholm School of Economics; David J. Johnson, Michigan State University; Kate M. Johnson, University of Southern California; Larissa G. Johnson, University of Birmingham; William J. Johnston, University of Chicago; Kai Jonas, University of Amsterdam; Jennifer A. Joy-Gaba, Virginia Commonwealth University; Heather Barry Kappes, London School of Economics and Political Science; Kim Kelso, Adams State University; Mallory C. Kidwell, Center for Open Science; Seung Kyung Kim, Stanford University; Matthew Kirkhart, Loyola University Maryland; Bennett Kleinberg, University College London, University of Amsterdam; Goran Knežević, University of Belgrade; Franziska Maria Kolorz, Radboud University Nijmegen; Jolanda J. Kossakowski, University of Amsterdam;

Robert Wilhelm Krause, University of Nijmegen; Job Krijnen, Tillburg University; Tim Kuhlmann, University of Konstanz; Yoram K. Kunkels, University of Amsterdam; Megan M. Kyc, Lock Haven University; Calvin K. Lai, University of Virginia; Aamir Laique, Saratoga, CA; Daniel Lakens, Eindhoven University of Technology; Kristin A. Lane, Bard College; Bethany Lassetter, University of Iowa; Ljiljana B. Lazarević, University of Belgrade; Etienne P. LeBel, Western University; Key Jung Lee, Stanford University; Minha Lee, University of Virginia; Kristi Lemm, Western Washington University; Carmel A. Levitan, Occidental College; Melissa Lewis, Reed College; Lin Lin, University of Hong Kong; Stephanie Lin, Stanford University; Matthias Lippold, University of Göttingen; Darren Loureiro, University of Maryland; Ilse Luteijn, Radboud University Nijmegen; Sean Mackinnon, Dalhousie University; Heather N. Mainard, Georgia Gwinnett College; Denise C. Marigold, Renison University College at University of Waterloo; Daniel P. Martin, University of Virginia; Tylar Martinez, Adams State University; E.J. Masicampo, Wake Forest University; Josh Matacotta, California State University, Fullerton; Maya Mathur, Stanford University; Michael May, Max Planck Institute for Research on Collective Goods, University of Bonn; Nicole Mechin, University of Alabama; Pranjal Mehta, University of Oregon; Johannes Meixner, Humboldt University of Berlin, University of Potsdam; Alissa Melinger, University of Dundee; Jeremy K. Miller, Willamette University; Mallorie Miller, Mississippi State University; Katherine Moore, Elmhurst College, Arcadia University; Marcus Möschl, Technische Universität Dresden; Matt Motyl, University of Illinios at Chicago; Stephanie M. Müller, University of Erfurt; Marcus Munafo, University of Bristol; Koen I. Neijenhuijs, Radboud University Nijmegen; Taylor Nervi, Ashland University; Gandalf Nicolas, William and Mary; Gustav Nilsonne, Stockholm University, Karolinska Institute; Brian A. Nosek, University of Virginia, Center for Open Science; Michèle B. Nuijten, Tilburg University; Catherine Olsson, New York University, Massachusetts Institute of Technology; Colleen Osborne, University of Virginia; Lutz Ostkamp, University of Osnabrück; Misha Pavel, Northeastern University; Ian S. Penton-Voak, University of Bristol; Olivia Perna, Ashland University; Cyril Pernet, The University of Edinburgh; Marco Perugini, University of Milan-Bicocca; R. Nathan Pipitone, Adams State University; Michael Pitts, Reed College; Franziska Plessow, Harvard Medical School, Technische Universität Dresden; Jason M. Prenoveau, Loyola University Maryland; Kate A. Ratliff, University of Florida; Rima-Maria Rahal, University of Amsterdam; David Reinhard, University of Virginia; Frank Renkewitz, University of Erfurt; Ashley A. Ricker, University of California, Riverside; Anastasia Rigney, University of Texas, Austin; Andrew M. Rivers, University of California, Davis; Mark Roebke, Wright State University; Abraham M. Rutchick, California State University, Northridge; Robert S. Ryan, Kutztown University of Pennsylvania; Onur Sahin, University of Amsterdam; Anondah Saide, University of California, Riverside; Gillian M. Sandstrom, University of British Columbia; David Santos, Universidad Autónoma de Madrid, IE Business School; Rebecca Saxe, Massachusetts Institute of Technology; René Schlegelmilch, University of Erfurt, Max Planck Institute for Research on Collective Goods; Kathleen Schmidt, University of Virginia's College at Wise; Sabine Scholz, University of Groningen; Larissa Seibel, Radboud University Nijmegen; Dylan Faulkner Selterman, University of Maryland; Samuel Shaki, Ariel University; William B. Simpson, University of Virginia; H. Colleen Sinclair, Mississippi State University; Jeanine L. M. Skorinko, Worchester Polytechnic Institute; Agnieszka Slowik, University of Vienna; Joel S. Snyder, University of Nevada, Las Vegas; Courtney Soderberg, Center for Open Science; Carina Sonnleitner, University of Vienna; Nick Spencer, Adams State University; Jeffrey R. Spies, Center for Open Science; Sara Steegen, University of Leuven; Stefan Stieger, University of Konstanz; Nina Strohminger, Duke University; Gavin B. Sullivan, Centre for Research in Psychology, Behaviour and Achievement, Coventry University; Thomas Talhelm, University of Virginia; Megan Tapia, Adams State University; Anniek te Dorsthorst, Radboud University Nijmegen; Manuela Thomae, University of Winchester, The Open University; Sarah L. Thomas, University of Virginia; Pia Tio, University of Amsterdam; Frits Traets, University of Leuven; Steve Tsang, City University of Hong Kong; Francis Tuerlinckx, University of Leuven; Paul Turchan, Jacksonville University; Milan Valášek, University of Edinburgh; Anna E. van 't Veer, Tilburg University, TIBER (Tilburg Institute for Behavioral Economics Research); Robbie Van Aert, Tilburg University; Marcel van Assen, Tilburg University; Riet van Bork, University of Amsterdam; Mathijs van de Ven, Radboud University Nijmegen; Don van den Bergh, University of Amsterdam; Marije van der Hulst, Radboud University Nijmegen; Roel van Dooren, Radboud University Nijmegen; Johnny van Doorn, University of Leuven; Daan R. van Renswoude, University of Amsterdam; Hedderik van Rijn, University of Groningen; Wolf Vanpaemel, University of Leuven; Alejandro Vásquez Echeverría, Universidad de la República Uruguay; Melissa Vazquez, Georgia Gwinnett College; Natalia

A core principle of scientific progress is reproducibility (*1-6*). Scientific claims should not gain credence because of the status or authority of their originator, but by the repeatability of their supporting evidence. Scientists attempt to transparently describe the methodology and resulting evidence used to support the claim. Other scientists agree or disagree whether the evidence supports the claim, citing theoretical or methodological reasons, or by collecting new evidence. Such debates are rendered meaningless, however, if the evidence being debated is not reproducible.

Even research of exemplary quality may have irreproducible empirical findings because of random or systematic error. Direct replication is the attempt to recreate the conditions believed sufficient for obtaining a previously observed finding (*7, 8*), and is the means of establishing reproducibility with new data. A direct replication may not obtain the original result for a variety of reasons: known or unknown differences between the replication and original study may moderate the size of an observed effect, the original result could have been a false positive, or the replication could produce a false negative. False positives and false negatives provide misleading information about effects; and, failure to identify the necessary and sufficient conditions to reproduce an effect indicates an incomplete theoretical understanding of the finding. Direct replication provides the opportunity to assess and improve reproducibility.

There is plenty of concern (*9-13*), but little direct evidence about the rate and predictors of reproducibility. What evidence does exist suggests that reproducibility is lower than desired or anticipated. Using Bayesian reasoning, Ioannidis (*9*) estimated that

---

Velez, Stanford University; Marieke Vermue, Radboud University Nijmegen; Mark Verschoor, Tilburg University; Michelangelo Vianello, University of Padua; Martin Voracek, University of Vienna; Gina Vuu, University of Virginia; Eric-Jan Wagenmakers, University of Amsterdam; Joanneke Weerdmeester, Radboud University Nijmegen; Ashlee Welsh, Adams State University; Erin C. Westgate, University of Virginia; Joeri Wissink, Tilburg University; Michael Wood, University of Winchester; Andy Woods, University of Oxford, Bristol University; Emily Wright, Adams State University; Sining Wu, Mississippi State University; Marcel Zeelenberg, Tillburg University; Kellylynn Zuni, Adams State University

publishing and analytic practices make it likely that more than half of research results are false, and therefore irreproducible. In cell biology, two industrial laboratories reported success replicating the original results of landmark findings in only 11% and 25% of the attempted cases (*10, 11*). These numbers are stunning, but they are also difficult to interpret because no details are available about the studies, methodology, or results. With no transparency, the reasons for low reproducibility cannot be evaluated.

Other investigations point to practices and incentives that may inflate the likelihood of obtaining false positive results in particular, or irreproducible results more generally. Potentially problematic practices include selective reporting, selective analysis, and insufficient specification of the conditions necessary or sufficient to obtain the results (*12-23*). In short, there are many informed hypotheses, but little direct evidence, about the reproducibility of published research. We were inspired to redress this gap. In this article, we report a large-scale, collaborative effort to obtain an initial estimate of the reproducibility of psychological science.

Starting in November 2011, we constructed a protocol for selecting and conducting high-powered, high-quality replications of a sample of published studies in psychology to maximize quality and generalizability (*24*). The project was hosted at a project management site (Open Science Framework; http://osf.io/) where collaborators joined the project, selected a study for replication from the available studies in the sampling frame, and were guided through the replication protocol. The replication protocol articulated the process of selecting the study and key effect from the available articles, contacting the original authors for study materials, preparing a study protocol and analysis plan, obtaining review of the protocol by the original authors and other members within the present project, registering the protocol publicly, conducting the replication, writing the final report, and auditing the process and

analysis for quality control (see *25*, and Supplementary Information for method details). Project coordinators facilitated each step of the process and maintained the protocol and project resources. Replication materials and data were required to be archived publicly to maximize transparency and accountability of the project team to the scientific community (https://osf.io/ezcuj/).

To explore correlates of reproducibility we assessed characteristics of the original study such as the publishing journal; original effect size, *p*-value, and sample size; experience and expertise of the original research team; rated importance of the effect as indicated by factors such as the citation impact of the article; and, rated surprisingness of the effect. We also assessed characteristics of the replication such as statistical power and sample size, experience and expertise of the replication team, independently assessed challenge of conducting an effective replication, and self-assessed quality of the replication effort.

In total, 100 replications were completed by 270 contributing authors. There were many different experimental designs and analysis strategies in the original research. Through consultation with original authors, obtaining original materials, and internal review, replications maintained high fidelity to the original designs. Analytic strategies converted results to a common effect size metric (*r*, correlation coefficient) and confidence intervals. The unit(s) of analysis for inferences about reproducibility were the original and replication studies. The resulting open dataset provides an initial estimate of the reproducibility of psychological science and correlational evidence to support development of hypotheses about the causes of reproducibility.

Results

Evaluating Reproducibility

There is no singular standard for evaluating replication success (*25*). Here, we

present five indicators of reproducibility—all of which contribute information about the cumulative evidence for the relations between the replication and original finding and the combined knowledge about the effect, and were correlated with one another (*r*'s range from .21 to .98, median = .57). Results are summarized in Table 1.

Evaluating replication effect against null hypothesis of no effect. A straightforward method for evaluating replication is to test whether the replication shows a statistically significant effect ($p < .05$) with the same direction as the original study. This dichotomous vote counting method is intuitively appealing and consistent with common heuristics used to decide if original studies "worked". 97 of 100 (97%) effects from original studies were positive results (4 had *p*-values falling a bit short of the .05 criterion, *p*'s = .0508, .0514, .0516, .0567, but all of these were interpreted as positive effects). Based just on the average replication power of the 97 original, significant effects (*M* = .92, *Mdn* = .95), we would expect approximately 89 positive results in the replications, however there were just 35 (36.1%; 95% CI = [26.6%, 46.2%]), a significant reduction (McNemar test, $X^2(1) = 59.5$, $p < .001$).

A key weakness of this method is that it treats the .05 threshold as a bright-line criterion between replication success and failure (*26*). It could be that many of the replications fell just short of the .05 criterion. The left panel of Figure 1 shows the density plots of *p*-values for original studies (Mean *p*-value = .028) and replications (Mean *p*-value = .302). The 64 non-significant *p*-values for replications were distributed widely. However, this distribution deviated slightly from uniform suggesting that at least one replication could be a false negative ($\chi^2(128) = 155.83$, $p = .048$). The wide distribution of *p*-values suggests against insufficient power as an overall explanation for failures to replicate. Figure 2 shows a scatterplot of original compared with replication study *p*-values.

Evaluating replication effect against original effect size. A complementary method for evaluating replication is to test whether the original effect size is within the 95% confidence interval of the effect size estimate from the replication. In other words, is the replication effect significantly different than the original estimate? For the subset of 74 studies in which the standard error of the correlation could be computed, 31 of the 74 (41.9%) of the replication confidence intervals contained the original effect size (significantly lower than the expected value of 78.5%, $p < .001$, see SI). For 21 studies using other test statistics ($F[df_1 > 1, df_2]$ and X^2), 66.7% of confidence intervals contained the effect size of the original study. Overall, this suggests a 47.4% replication success rate on this criterion.

This method addresses the weakness of the first test—a replication in the same direction and a *p*-value of .06 may not be significantly different from the original result. However, the method will also indicate that a replication "fails" when the direction of the effect is the same but the replication effect size is significantly smaller than the original effect size (*27*). Also, the replication "succeeds" when the result is near zero but not estimated with sufficiently high precision to be distinguished from the original effect size.

Comparing original and replication effect sizes. Comparing the magnitude of the original and replication effect sizes avoids special emphasis on *p*-values. Overall, original study effect sizes ($M = .396$, $SD = .193$) were reliably larger than replication effect sizes ($M = .198$, $SD = .255$), *Wilcoxon's W* = 7132, $p < .001$. Of the 97 studies for which an effect size in both the original and replication study could be calculated (*28*), 82 showed a stronger effect size in the original study (82.8%; $p < .001$, binomial test; see right panel of Figure 1). Original and replication effect sizes were positively correlated (Spearman's $r = .51$, $p < .001$). Figure 3 presents a scatterplot of the original and replication effect sizes.

Combining original and replication effect sizes for cumulative evidence. The

disadvantage of the descriptive comparison of effect sizes is that it does not provide information about the precision of either estimate, or resolution of the cumulative evidence for the effect. This is often addressed by computing a meta-analytic estimate of the effect sizes by combining the original and replication studies (*26*). This approach weights each study based on sample size, and uses these weighted estimates of effect size to estimate cumulative evidence and precision of the effect. Using a fixed-effect model, 52 of the 74 (70.3%) effects for which a meta-analytic estimate could be computed had 95% confidence intervals that did not include 0.

An important qualification about this result is the possibility that the original studies have inflated effect sizes due to publication, selection, reporting, or other biases (*9, 12-23*). In a discipline with low powered research designs and an emphasis on positive results for publication, effect sizes will be systematically overestimated in the published literature. There is no publication bias in the replication studies because all results are reported. Also, there are no selection or reporting biases because they were confirmatory tests based on pre-analysis plans. This maximizes the interpretability of the replication *p*-values and effect estimates. If publication, selection, and reporting biases completely explain the effect differences, then the replication estimates would be a better estimate of the effect size than the meta-analytic and original results. However, to the extent that there are other influences, such as moderation by sample, setting, or quality of replication, the relative bias influencing original and replication effect size estimation is unknown.

Subjective assessment of "Did it replicate?" In addition to the quantitative assessments of replication and effect estimation, replication teams provided a subjective assessment of replication success. Subjective assessments of replication success were very similar significance testing results (38 of 100 successful replications), including evaluating

"success" for two null replications when the original study reported a null result, and "failure" for a $p < .05$ replication when the original result was a null.

Correlates of Reproducibility. Table 1 summarizes the overall replication evidence across the criteria described above, and then separately by journal/discipline. Considering significance testing, reproducibility was stronger in studies and journals representing cognitive psychology than social psychology topics. For example, combining across journals, 14 of 53 (26%) of social psychology effects replicated by the $p < .05$ criterion, whereas 20 of 38 (53%) of cognitive psychology effects did so. Simultaneously, all journals and disciplines showed substantial and similar ($X^2(4) = 5.91$, $p = .21$) decline in effect size in the replications compared to the original studies. The difference in significance testing results between fields appears to be partly a function of weaker original effects in social psychology studies, particularly in JPSP, and perhaps greater frequency of high-powered within-subjects manipulations and repeated measurement designs in cognitive psychology as suggested by high power despite small participant samples. Further, the type of test was associated with replication success. Among original, significant effects, 23 of the 50 (46%) that tested main or simple effects replicated at $p < .05$, but just 8 of the 37 (22%) that tested interaction effects did.

Table 2 (and Tables S4 and S5) provide correlations between reproducibility indicators and characteristics of replication and original studies. A negative correlation with the original study *p*-value indicates that the initial strength of evidence is predictive of reproducibility. For example, 26 of 63 (41%) original studies with $p < .02$ achieved $p < .05$ in the replication, whereas 6 of 23 (26%) that had a *p*-value between $.02 < p < .04$, and 2 of 11 (18%) that had a *p*-value > .04 did so (see Figure 2). Almost ⅔ (20 of 32, 63%) of original studies with $p < .001$ had a significant *p*-value in the replication.

Larger original effect sizes were associated with greater likelihood of achieving $p < .05$ ($r = .304$) *and* a greater decline in effect size observed in the replication ($r = .310$). Moreover, replication success ($p < .05$) was related to replication power ($r = .374$) but not with decline in effect size in the replication ($r = -.065$). Comparing effect sizes across indicators, there are hints that the surprisingness of the original effect and the challenge of conducting the replication are related to replication success, but these effects were relatively small and not observed consistently across indicators. Finally, there was little evidence that perceived importance of the effect, experience and expertise of the original or replication teams, and self-assessed quality of the replication accounted for meaningful variation in reproducibility across indicators. In sum, replication success was more consistently related with the original strength of evidence (e.g., original *p*-value, effect size, and effect tested) than with characteristics of the teams and implementation of the replication (e.g., expertise, quality, challenge of conducting study; see also SI Tables S4 and S5).

Discussion

We conducted a large-scale, collaborative effort to reproduce 100 findings from the published psychology literature. Replication effects ($M = .198$, $SD = .255$) were half the magnitude of original effects ($M = .396$, $SD = .196$) representing a substantial decline effect (*29*). Thirty-six percent of replications had significant results ($p<.05$) compared to 97% of original studies; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 38% of effects were subjectively rated to have replicated the original result; and, with the assumption of no publication bias in original results, 70% of the meta-analytic combinations of original and replication results were significant effects. No single indicator sufficiently describes replication success, and these indicators are not the only ways to evaluate reproducibility. Nonetheless, collectively, these results offer a clear

conclusion: a large portion of replications did not reproduce evidence supporting the original results despite using high-powered designs, materials provided by the original authors, and review in advance for methodological fidelity. Moreover, correlational evidence is consistent with the conclusion that variation in the strength of initial evidence was more predictive of replication success than variation in the characteristics of the teams and conducting of the replication. The latter factors certainly can influence replication success, the evidence is just that they did not do so here. Other investigators may develop alternative indicators to explore further the role of expertise and quality in reproducibility on this open dataset.

Insights on Reproducibility

It is too easy to conclude that successful replication means that the theoretical understanding of the original finding is correct. However, direct replication only provides evidence for the reliability of a result. If there are alternative explanations for the original finding, those alternatives could likewise account for the replication. Understanding is achieved through multiple, diverse investigations that provide converging support for a theoretical interpretation and rule out alternative explanations.

It is also too easy to conclude that a failure to replicate a result means that the original evidence was a false positive. There is good evidence that low power of research designs combined with publication bias favoring positive results over negative results produces a literature with upwardly biased effect sizes (*14, 16, 30, 31*). This anticipates that replication effect sizes would be smaller than original studies on a routine basis—not because of differences in implementation but because the original study effect sizes are affected by publication and reporting bias and the replications are not. The results showed a decline effect consistent with this expectation. Most replication effects were smaller than original results, and reproducibility success was correlated with indicators of the strength of initial

evidence such as lower original *p*-values and larger effect sizes. This suggests publication, selection, and reporting biases as plausible explanations for the difference between original and replication effects. The replication studies have none of these biases because replication pre-registration and pre-analysis plans ensured confirmatory tests and reporting of all results.

Finally, besides revealing false positives, replications could fail if the replication methodology differs from the original in ways that interfere with observing the effect. We conducted replications designed to minimize *a priori* reasons to expect a different result by using original materials, engaging original authors for review of the designs, and with internal review. Nonetheless, none of these efforts guarantee that unanticipated factors in the sample, setting, or procedure altered the observed effect magnitudes (*32*). The observed variation in replication and original results may reduce certainty about the statistical inferences of the original study, but it also provides opportunity for theoretical innovation to explain differing outcomes, to be followed by research to test those hypothesized explanations.

Implications and Limitations

The present study provides the first open, systematic evidence of reproducibility from a sample of studies in psychology. We sought to maximize generalizability of the results with a structured process for selecting studies for replication. However, it is unknown the extent to which these findings extend to the rest of psychology subdisciplines, or to other disciplines. In the sampling frame itself, not all studies were replicated. For example, more resource intensive studies were less likely to be included than less resource intensive studies. While study selection bias was reduced by the sampling frame and selection strategy (see SI), the impact of selection bias is unknown.

We investigated the reproducibility rate of psychology, not because there is something special about psychology, but because it is our discipline. Concerns about reproducibility are widespread across disciplines (*9-21*). Reproducibility is not well-understood because the incentives for individual scientists prioritize novelty over replication. If nothing else, this project demonstrates that it is possible to conduct a large-scale examination of reproducibility despite the incentive barriers. Perhaps this, and the related Many Labs replication projects (*32*), will spur similar efforts across disciplines such as the ongoing effort in cancer biology (*33*). Cumulative evidence across fields would yield insight about common and unique challenges, and may cross-fertilize strategies to improve reproducibility.

Because reproducibility is a hallmark of credible scientific evidence, it is tempting to think that maximum reproducibility of original results is important from the onset of a line of inquiry through its maturation. This is a mistake. If initial ideas were always correct, then there would be hardly a reason to conduct research in the first place. A healthy discipline will have many false starts as it confronts the limits of present understanding.

Innovation is the engine of discovery and is vital for a productive, effective scientific enterprise. However, innovative ideas become old news fast. Journal reviewers and editors may dismiss a new test of a published idea as unoriginal. The claim that “we already know this” belies the uncertainty of scientific evidence. Deciding the ideal balance of resourcing innovation versus verification is a question of research efficiency. How can we maximize the rate of research progress? Innovation points out paths that are possible; replication points out paths that are likely; progress relies on both. The ideal balance is a topic for investigation itself. Scientific incentives - funding, publication, awards - can be tuned to encourage an optimal balance in the collective effort of discovery (TOP Guidelines, this issue; *34*).

Progress occurs when existing expectations are violated, and a surprising result spurs a new investigation. Replication can increase certainty when findings are reproduced, and promote innovation when they are not. This project provides accumulating evidence for many findings in psychological research, and suggests that there is still more work to do to verify whether we know what we think we know.

Conclusion

Following this intensive effort to reproduce a sample of published psychological findings, how many of the effects can we confirm are true? Zero. And, how many of the effects can we confirm are false? Zero. Is this a limitation of the project design? No. It is the reality of doing science, even if it is not appreciated in daily practice. Humans desire certainty and science infrequently provides it. As much as we might wish it to be otherwise, a single study almost never provides definitive resolution for or against an effect and its explanation. The original studies examined here offered tentative evidence, the replications we conducted offered additional, confirmatory evidence. In some cases, the replications increase confidence in the reliability of the original results; in other cases, the replications suggest that more investigation is needed to establish credibility of the original findings. Scientific progress is an accumulating process of uncertainty reduction that can only succeed if science itself remains the greatest skeptic of its explanatory claims.

The present results suggest that there is room to improve reproducibility in psychological science. Any temptation to interpret these results as a defeat for psychology, or science more generally, must contend with the fact that this project demonstrates science behaving as it should. Hypotheses abound that the present culture in science may be negatively affecting the reproducibility of findings. An ideological response would discount the arguments, discredit the sources, and proceed merrily along. The scientific process is not

ideological. Science does not always provide comfort for what we wish to be; it confronts us with what is. Moreover, as illustrated by the TOP Guidelines (this issue), researchers are taking action already to improve the quality and credibility of the scientific literature.

We conducted this project because we care deeply about the health of our discipline, and believe in its promise for accumulating knowledge about human behavior that can advance the quality of the human condition. Reproducibility is central to that aim. Accumulating evidence is the scientific community's method of self-correction and it is the best available option for achieving that ultimate goal - truth.

References

1. C. Hempel, Maximal specificity and lawlikeness in probabilistic explanation. *Philos. Sci*.**35**, 116–133 (1968).

2. C. Hempel, P. Oppenheim, Studies in the logic of explanation. *Philos. Sci*.**15**, 135–175 (1948).

3. I. Lakatos, in *Criticism and the Growth of Knowledge*, I. Lakatos, A. Musgrave, Eds. (Cambridge Univ. Press, London, 1970) pp. 170-196.

4. P. E. Meehl, Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychol. Inq*.**1**, 108–141 (1990).

5. J. Platt, Strong inference. *Science***146**, 347–353 (1964).

6. W. C. Salmon, in *Introduction to the Philosophy of Science,* M. H. Salmon Ed. (Hackett Publishing Company, Inc., Indianapolis, 1999) pp. 7-41.

7. B. A. Nosek, D. Lakens, Registered reports: A method to increase the credibility of published results. *Soc. Psychol.***45**, 137-141 (2014).

8. S. Schmidt, Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Rev. Gen. Psychol.***13***,* 90-100 (2009).

9. J. P. A. Ioannidis, Why most published research findings are false. *PLoS Med.***2***,* e124 (2005), doi: 10.1371/journal.pmed.0020124.

10. C. G. Begley, L. M. Ellis, Raise standards for preclinical cancer research. *Nature***483**, 531-533 (2012).

11. F. Prinz, T. Schlange, K. Asadullah, Believe it or not: How much can we rely on published data on potential drug targets? *Nat. Rev. Drug Disc.***10**, 712-713 (2011

12. M. McNutt, Reproducibility. *Science*,**343**, 229 (2014).

13. H. Pashler, E-J. Wagenmakers, Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspect. Psychol. Sci.***7**, 528-530 (2012).

14. K. S. Button, *et al*., Power failure: Why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.***14***,* 1-12 (2013).

15. D. Fanelli, “Positive” results increase down the hierarchy of the Sciences. *PLoS One**5***, e10068 (2010), doi: 10.1371/journal.pone.0010068 .

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

17. G. S. Howard, *et al*., Do research literatures give correct answers? *Rev. Gen. Psychol.***13**, 116-121 (2009).

18. J. P. A. Ioannidis, M. R. Munafo, P. Fusar-Poli, B. A. Nosek, S. P. David, Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends Cogn. Sci.***18**, 235-241 (2014).

19. L. John, G. Loewenstein, D. Prelec, Measuring the prevalence of questionable research practices with incentives for truth-telling*. Psychol. Sci.***23**, 524-532 (2012).

20. B. A. Nosek, J. R. Spies, M. Motyl, Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspect. Psychol. Sci.***7***,* 615-631 (2012).

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

22. P. Rozin, What kind of empirical research should we publish, fund, and reward?: A different perspective. *Perspect. Psychol. Sci.***4**, 435-439 (2009).

23. J. P. Simmons, L. D. Nelson, U. Simonsohn, False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.***22***,* 1359-1366 (2011).

24. Open Science Collaboration, An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspect. Psychol. Sci.***7***,* 657-660 (2012).

25. Open Science Collaboration, in *Implementing Reproducible Computational Research (A Volume in The R Series)*, V. Stodden, F. Leisch, R. Peng, Eds. (Taylor & Francis, New York, 2014) pp. 299-323.

26. S. L. Braver, F. J. Thoemmes, R. Rosenthal, Continuously cumulating meta-analysis and replicability. *Perspect. Psychol. Sci.***9**, 333-342 (2014).

27. U. Simonsohn, Small telescopes: Detectability and the evaluation of replication results. *Psychol. Sci.* (2015), doi: 10.1177/0956797614567341.

28. D. Lakens, Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Front. Psychol.***4**, 863 (2013), doi: 10.3389/fpsyg.2013.00863.

29. J. Lehrer, The truth wears off: Is there something wrong with the scientific method? *The New Yorker*, 52-57 (2010).

30. J. Cohen, The statistical power of abnormal-social psychological research: A review. *J. Abnorm. Soc. Psychol.***65**, 145-153 (1962).

31. T. D. Sterling, Publication decisions and their possible effects on inferences drawn from tests of significance-or vice versa. *J. Am. Stat. Assoc.***54**, 30-34 (1959).

32. R. Klein, *et al*., Investigating variation in replicability: A “many labs” replication project. *Soc. Psychol.***45**, 142-152 (2014).

33. T. Errington, *et al*., An open investigation of the reproducibility of cancer biology research. *eLife***3**, e04333 (2014), doi: 10.7554/eLife.04333.

34. J. K. Hartshorne, A. Schachner, Tracking replicability as a method of post-publication open evaluation. *Front. Comput. Neurosci* (2012), doi:10.3389/fncom.2012.00008

35. P. Bressan, D. Stranieri, The best men are (not always) already taken: Female preference for single versus attached males depends on conception risk. *Psychol. Sci.***19***,* 145-151 (2008).

36. D. Albarracín, *et al*., Increasing and decreasing motor and cognitive output: A model of general action and inaction goals. *J. Pers. Soc. Psychol*.**95**, 510-523 (2008).

37. R. Rosenthal, K. L. Fode, The effect of experimenter bias on the performance of the albino rat. *Behav. Sci*.**8**, 183-189 (1963).

38. R. A. Fisher, Theory of statistical estimation. *Math. Pro. Camb. Phil. Soc.***22**, 700-725 (1925).

39. G. Cumming, The new statistics: why and how. *Psychol. Sci.***25**, 7-29 (2013).

Supplementary Materials
www.sciencemag.org
Materials and Methods
Figs. S1-S7
Tables S1-S5
References (35-39)

Acknowledgments

In addition to the co-authors of this manuscript, there were many volunteers that contributed to project success. We thank David Acup, John Anderson, Stefano Anzellotti, Robson Araujo, Jack D. Arnal, Timothy Bates, Ruairidh Battleday, Robert Bauchwitz, Mira Bernstein, Ben Blohowiak, Marilisa Boffo, Emile Bruneau, Benjamin Chabot-Hanowell, Joel Chan, Phuonguyen Chu, Anna Dalla Rosa, Benjamin Deen, Philip DiGiacomo, Canay Dogulu, Nicholas Dufour, Cailey Fitzgerald, Adam Foote, Alexander Garcia, Emmanuel Garcia, Chantal Gautreau, Laura Germine, Tripat Gill, Lisa Goldberg, Stephen D. Goldinger, Hyowon Gweon, Dibora Haile, Kathleen Hart, Frederik Hjorth, Jillian Hoenig, Åse Innes-Ker, Brenda Jansen, Radka Jersakova, Yun Jie, Zsuzsa Kaldy, Wai Keen Vong, Avril Kenney, John Kingston, Jorie Koster-Hale, Ann Lam, Richard LeDonne, Daniel Lumian, Emily Luong, Sally Man-pui, Jessica Martin, Amy Mauk, Todd McElroy, Kateri McRae, Tyler Miller, Katharina Moser, Michael Mullarkey, Alisa R. Munoz, Joanne Ong, Colleen Parks, Debra Sue Pate, David Patron, Helena J. M. Pennings, Michael Penuliar, Angela Pfammatter, Joseph Phillip Shanoltz, Emily Stevenson, Emily Pichler, Henriette Raudszus, Hilary Richardson, Nina Rothstein, Thomas Scherndl, Sheree Schrager, Shraddha Shah, Yan Shan Tai, Amy Skerry, Mia Steinberg, Jonathan Stoeterau, Helen Tibboel, Anastasia Tooley, Alexa Tullett, Christian Vaccaro, Evie Vergauwe, Aimi Watanabe, Ian Weiss, Mark H. White II, Paul Whitehead, Catherine Widmann, David K. Williams, Krista M. Williams, and Hu Yi.

Also, we thank the authors of the original research that was the subject of replication in this project. These authors were generous with their time, materials, and advice for improving the quality of each replication and identifying the strengths and limits of the outcomes.

Supplemental Information for
Estimating the Reproducibility of Psychological Science

Open Science Collaboration

Table of Contents

Method

Two articles have been published on the methodology of the Reproducibility Project: Psychology.

1. Open Science Collaboration, An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspect. Psychol. Sci.***7***,* 657-660 (2012).
2. Open Science Collaboration, in *Implementing Reproducible Computational Research (A Volume in The R Series)*, V. Stodden, F. Leisch, R. Peng, Eds. (Taylor & Francis, New York, 2014) pp. 299-323.

The first introduced the project aims and basic design. The second provided detail on the methodology and mechanisms for maintaining standards and quality control. The methods sections below summarize the key aspects of the methodology and provide additional information, particularly concerning the latter stages of the project that were not addressed in the prior articles.

Sampling Frame and Study Selection

We pursued a quasi-random sample by defining the sampling frame as the 2008 articles of three important psychology journals—*Psychological Science (PSCI)*, *Journal of Personality and Social Psychology*, and *Journal of Experimental Psychology: Learning, Memory, and Cognition*. The first is a premiere outlet for all psychological research, the second and third are leading disciplinary-specific journals for social psychology and cognitive psychology respectively (see Open Science Collaboration, 2012 for more information).

Teams selected articles matching their research interests and expertise from an initial pool of the first 30 articles from each journal in chronological order. As articles were claimed, additional articles were made available in groups of 10. This approach balanced competing goals: minimizing selection bias by having only a small set of articles available at a time, and matching studies with replication teams' interests, resources, and expertise.

In total, there were 488 articles in the 2008 issues of the three journals. 158 of these (32%) became eligible for selection during the project period, between November 2011 and December 2014. From those, 111 (70%) articles were selected by a replication team, producing 113 replications[2]. And, 100 of those (88%) replications were completed by the project deadline for inclusion in this aggregate report. Of the 47 articles from the eligible sample that were not claimed, 6 (13%) had been deemed infeasible to replicate by a group of volunteers because of time, resources, instrumentation, dependence on historical events, or hard to access samples. The remaining 41 (87%) were eligible but were not claimed. Eligible

[2] Two of the articles available for replication were replicated twice (*35, 36*). The first (*35*), was replicated in an in lab setting, and online as a secondary replication. The second, experiment 7 of Albarracín et al. (2008) was replicated in a lab setting and a secondary replication of experiment 5 was conducted online. These two supplementary replications bring the total number of replications pursued to 113 and total completed to 100.

but unclaimed studies often required specialized resources or knowledge—access to autistic participants, macaque monkeys, or eye tracking machines, for example.

By default, the last experiment reported in each article would be the subject of replication. Deviations from selecting the last experiment were made occasionally based on feasibility or recommendations of the original authors. Justification for deviations were report in the replication reports, made available on the Open Science Framework (http://osf.io/ezcuj). In total, 84 of the 100 completed replications (84%) were of the last study in the article. On average, the to-be-replicated articles contained 2.99 studies (SD = 1.78) with the following distribution: 24 single study, 24 two studies, 18 three studies, 13 four studies, 12 five studies, 9 six or more studies. All following summary statistics refer to the 100 completed replications.

For the purposes of aggregating results across studies to estimate reproducibility, a key result from the selected experiment was identified as the focus of replication. This effect was identified prior to data-collection or analyses and was presented to the original authors as part of an initial report. In the individual replication reports and subjective assessments of replication outcomes, more than a single result may be examined, but only the result of the key effect was considered in the present analyses.

Replication Teams

RPP was introduced publicly as a crowdsourcing research project in November 2011. Interested researchers were invited to get involved to design the project, conduct a replication, or provide other kinds of research support such as coding articles. A total of 259 individuals contributed sufficiently to earn co-authorship on this report.

Of the 100 replication completed, 85 unique senior members were identified—several of whom led multiple replications. Among those senior members, 72 had a PhD or equivalent, 9 had a master's degree or equivalent, 1 had some graduate school, and 3 had or were near completing a bachelor's degree or equivalent. By occupation, 62 were faculty members or equivalent, 8 were post-docs, 13 were graduate students, 1 was an undergraduate student, and 1 was a private sector researcher. By domain, 36 identified social psychology as their primary domain, 22 identified cognitive psychology, 6 identified quantitative psychology, and 21 identified other domains.

Replication Protocol

Sloppy or underpowered replication attempts would provide uninteresting reasons for irreproducibility. Replication teams followed an extensive protocol to maximize quality, clarity, and standardization of the replications. Full documentation of the protocol is available at https://osf.io/ru689/.

Power analysis. After identifying the key effect, power analyses estimated the sample sizes needed to achieve 80%, 90%, and 95% power to detect the originally reported effect size. Teams were required to propose a study design that would achieve at least 80% power and were encouraged to obtain higher power if feasible to do so. All protocols proposed 80% power or greater, however, after corrections to power analyses, three fell short

in their planning, with 56%, 69%, and 76% power. On average, 92% power was proposed (median = 95%). Three replication teams were unable to conduct power analyses based on available data—their method for planning sample size is detailed in their replication reports. Following data collection, 90 of the 97 achieved 80% or greater power to detect the original effect size. Post-hoc calculations showed an average of 92% power to detect an effect size equivalent to the original studies'. The median power was 95% and 57 had 95% power or better. Note that these power estimates do not account for the possibility that the published effect sizes are overestimated because of publication biases. Indeed, this is one of the potential challenges for reproducibility.

Obtaining or creating materials. Project coordinators or replication teams contacted original authors for study materials in order to maximize the consistency between the original and replication effort. Of the completed replications, 89 were able to obtain some or all of the original materials. In 8 cases, the original materials were not available, and in only 3 cases the original authors did not share materials or provide information about where the materials could be obtained. Replication teams prepared materials, adapting or creating them for the particular data collection context. If information available from the original report or author contacts was insufficient, teams noted deviations or inferences in their written protocols.

Writing study protocols. The protocols included a brief introduction explaining the main idea of the study, the key finding for replication, and any other essential information about the study. Then, they had a complete methods section describing the power analysis, sampling plan, procedure, materials, and analysis plan. Analysis plans included details of data exclusion rules, data cleaning, inclusion of covariates in the model, and the inferential test/model that would be used. Finally, the protocol listed known differences from the original study in sampling, setting, procedure, and analysis plan. The objective was to minimize differences that are expected to alter the effect, but report transparently about them to provide a means of identifying possible reasons for variation in observed effects, and to identify factors for establishing generalizability of the results when similar effects are obtained. All replication teams completed a study protocol in advance of data collection.

Replicators were encouraged to apply for funding for the replication to the Center for Open Science—a grants committee comprised of members of the collaboration reviewed study protocols and award requests.

Reviewing study protocols. The written protocols were shared with original authors for critique prior to initiating data collection. Also, protocols were reviewed by another member of the RPP team for quality assurance and consistency with the reporting template. Feedback from the original authors was incorporated into the study design. If the replication team could not address the feedback, the original author comments were included in the protocol so that readers could identify the *a priori* comments by original authors about the design. Replication teams recorded whether the original authors endorsed the design (69 replications), maintained concerns based on informed judgment/speculation (8 replications), maintained concerns based on published empirical evidence for constraints on the effect (3 replications), or did not respond (18 replications). Two replications that were conducted prior to the sharing policy being established did not receive feedback prior to data collection.

Uploading the study protocol. Once finalized, the protocol and shareable materials were posted publicly on the Open Science Framework (OSF; https://osf.io/ezcuj/) following a standard format. If the original author requested to keep materials private, replication teams noted this and indicated how to contact the original author to obtain the materials. After upload, the replication team could begin data collection.

Reporting. Following data collection, teams initiated report writing and data sharing. If there were any deviations from the registered protocol, teams noted those in the final report. Also, teams posted anonymized datasets and a codebook to the OSF project page. Teams conducted the planned data analysis from the protocol as a confirmatory analysis. Following completion of the confirmatory analysis phase, teams were encouraged to conduct follow-up exploratory analysis if they wished and report both—clearly distinguished—in their final report.

After writing the results section of the final report, teams added discussion with open-ended commentary about insights gained from exploratory analysis, an overall assessment of the outcome of the replication attempt, and discussion of any objections or challenges raised by the original authors' review of the protocol. At least one other RPP member then conducted a review of the final report to maximize consistency in reporting format, identify errors, and improve clarity. Following review, replication teams shared their report directly with the original authors and publicly on the OSF project page. If additional issues came up following posting of the report, teams could post a revision of the report. The OSF offers version control so all prior versions of posted reports can be retrieved in order to promote transparent review of edits and improvements.

Data Preparation

To prepare the aggregate data and final report, we conducted project-wide audit of all projects, materials, data, and reports. Every project was reviewed to ensure that the formatting and reports conformed to the template and availability standards and that the conducted analysis and code could be reproduced. A description of this review is available on the OSF (https://osf.io/xtine/). Moreover, 27 members of the team reproduced the analyses of every experiment using a standardized format in the R analysis package as an audit the replication analyses. A controller R script was created to re-generate the entire analysis of every experiment and recreate the master datafile. This R script, available at https://osf.io/fkmwg/, can be executed to reproduce the results of the individual studies. A comprehensive description of this re-analysis process is available on the OSF (https://osf.io/a2eyg/).

In parallel, a subset of the RPP team defined variables to include in an examination of correlates of reproducibility. These variables could represent characteristics of the original study, characteristics of the replication, or contextual characteristics. Through this process, a single master data file was created for analysis.

Measures and Moderators

We assessed features of the original study and replication as possible correlates of

reproducibility. The master data file was populated with dozens of variables that are available for analysis. To reduce the likelihood of false positives due to many tests, we aggregated some individual variables into summary indicators: experience and expertise of original team, experience and expertise of replication team, challenge of replication, self-assessed quality of replication, and importance of the effect. Exploratory analyses of the individual variables are available in supplementary information. We had little *a priori* justification to favor particular variables over others, so aggregates were created by standardizing (M = 0, SD = 1) the individual variables and then averaging to create a single index. In addition to the publishing journal and subdiscipline, potential moderators included 6 characteristics of the original study, and 5 characteristics of the replication.

Publishing journal and subdiscipline. Journals' differences in publishing practices may result in a selection bias that covaries with reproducibility. Articles from three journals were made available for selection: JPSP (n=59), JEP:LMC (n=40), and PSCI (n=68). PSCI publishes articles across sub-disciplines in psychology. From this pool of available studies, completed replications were selected and completed from JPSP (n=32), JEP:LMC (n=28), and PSCI (n=40), and were coded as belonging to cognitive (n=37), social (n=54), or another discipline of psychology (n=9). For analysis, all studies appearing in JPSP were treated as from social psychology, and all studies appearing in JEP:LMC were treated as from cognitive psychology. Reproducibility may vary by sub-discipline in psychology because of differing practices. For example, within-subjects' designs are more common in cognitive than social psychology and these designs have greater power to detect effects with the same number of participants.

Characteristics of Original Study

Original study effect size, *p*-value, and sample size. Qualities of the original statistical evidence may predict reproducibility. All else being equal, results with larger effect sizes and smaller p-values ought to be more reproducible than others. Also, larger sample sizes are a factor for increasing the precision of estimating effects; all else being equal, larger sample sizes should be associated with more reproducible results. A qualification of this expectation is that some study designs use very few participants and gain substantial power via repeated measurements.

Importance of the result. Some effects are more important than others. This variable was the aggregate of the citation impact of the original article and coder ratings of the extent to which the article was exciting and important. Effect importance could be a positive predictor of reproducibility because findings that have a strong impact on the field do so, in part, because they are reproducible and spur additional innovation. If they were not reproducible, then they may not have a strong impact on the field. On the other hand, exciting or important results are appealing because they advance an area of research, but they may be less reproducible than mundane results because true advances are difficult and infrequent, and theories and methodologies employed at the fringe of knowledge are often less refined or validated making them more difficult to reproduce.

Citation impact of original article. Project coordinators used Google Scholar data to

calculate the citation impact of the original article at the time of conducting the project analysis (March 2015).

Exciting/important effect. Coders independent from the replication teams reviewed the methodology for the replication studies and answered the following prompt: “To what extent is the key effect an exciting and important outcome?” To answer this question, at least three coders read the pre-data collection reports that the replication teams had created. These reports included a background on the topic, a description of the effect, a procedure, and analysis plan. Responses were provided on a scale from 1 = Not at all exciting and important, 2 = Slightly exciting and important, 3 = Somewhat exciting and important, 4 = Moderately exciting and important, 5 = Very exciting and important, 6 = Extremely exciting and important.

Surprising result. Counterintuitive results are appealing because they violate one’s priors, but they may be less reproducible if priors are reasonably well-tuned to reality. At least three coders independent from the replication teams reviewed the methodology for the replication studies and answered the following prompt: “To what extent is the key effect a surprising or counter-intuitive outcome?” Coders read the pre-data collection reports Responses were provided on a scale from 1 = Not at all surprising, 2 = Slightly surprising, 3 = Somewhat surprising, 4 = Moderately surprising, 5 = Very surprising, 6 = Extremely surprising.

Experience and expertise of original team. Higher quality teams may produce more reproducible results. Quality is multi-faceted and difficult to measure. In the present study, we aggregated four indicators of quality - the rated prestige of home institutions of the 1st and senior authors, and the citation impact of the 1st and senior authors. Other means of assessing quality could reveal results quite distinct from those obtained by these indicators.

Institution prestige of 1st author and senior author. Authors were coded as being 1st and most senior; their corresponding institutions were also recorded. The resulting list was presented to two samples (Mechanical Turk participants n = 108; Project team members n = 70) to rate institution prestige on a scale from 7 = never heard of this institution, 6 = not at all prestigious, 5 = slightly prestigious, 4 = moderately prestigious, 3 = very prestigious, 2 = extremely prestigious, 1 = one of the few most prestigious. Project team members were randomly assigned to rate institution prestige *in psychology (n = 33)* or *in general (n = 37)*. Correlations of prestige ratings among the three samples were very high (r’s range .849 to .938). As such, before standardizing, we averaged the three ratings for a composite institution prestige score.

Citation impact of 1st author and senior author. Project members used Google Scholar data to estimate the citation impact of first authors and senior authors. These indicators identified citation impact at the time of writing this report, not at the time the original research was conducted.

Characteristics of Replication

Replication power and sample size. All else equal, lower power and smaller sample tests ought to be less likely to reproduce results than higher power and larger sample tests.

The caveat above on sample size for original studies is the same as for replication studies. Replications were required to achieve at least 80% power based on the effect size of the original study. This narrows the range of power in replication tests to maximize likelihood of obtaining effects, but nonetheless offers a range that could be predictive of reproducibility. A qualification of this expectation is that power estimates are based on original effects. If publication bias or other biases produce exaggerated effect sizes in the original studies, then the power estimates would be less likely to provide predictive power for reproducibility.

Challenge of conducting replication. Reproducibility depends on effective implementation and execution of the research methodology. However, some methodologies are more challenging or prone to error and bias than others. As a consequence, variation in the challenges of conducting replications may be a predictor of reproducibility. This indicator includes coders assessments of expertise required, opportunity for experimenter expectations to influence outcomes, and opportunity for lack of diligence to influence outcomes.

Perceived expertise required. Reproducibility might be lower for study designs that require specialized expertise. Coders independent from the replication teams reviewed the methodology for the replication studies and answered the following prompt: "To what extent does the methodology of the study require specialized expertise to conduct effectively? [Note: This refers to data collection, *not* data analysis]" Responses were provided on a scale from 1 = no expertise required, 2 = slight expertise required, 3 = moderate expertise required, 4 = strong expertise required, 5 = extreme expertise required.

Perceived opportunity for expectancy biases. The expectations of the experimenter can influence study outcomes (*37*). Study designs that provide opportunity for researchers' beliefs to influence data collection may be more prone to reproducibility challenges than study designs that avoid opportunity for influence. Coders independent from the replication teams reviewed the methodology for the replication studies and answered the following prompt: "To what extent does the methodology of the study provide opportunity for the researchers' expectations about the effect to influence the results? (i.e., researchers belief that the effect will occur could elicit the effect, or researchers belief that the effect will not occur could eliminate the effect) [Note: This refers to data collection, *not* data analysis]." Responses were provided on a scale from 1 = No opportunity for researcher expectations to influence results, 2 = Slight opportunity for researcher expectations to influence results, 3 = Moderate opportunity for researcher expectations to influence results, 4 = Strong opportunity for researcher expectations to influence results, 5 = Extreme opportunity for researcher expectations to influence results.

Perceived opportunity for impact of lack of diligence. Studies may be less likely to be reproducible if they are highly reliant on experimenters' diligence to conduct the procedures effectively. Coders independent from the replication teams reviewed the methodology for the replication studies and answered the following prompt: "To what extent could the results be affected by lack of diligence by experimenters in collecting the data? [Note: This refers to data collection, not creating the materials]." Responses were provided on a scale from 1 = No opportunity for lack of diligence to affect the results, 2 = Slight opportunity for lack of diligence to affect the results, 3 = Moderate opportunity for lack of diligence to affect the results, 4 = Strong opportunity for lack of diligence to affect the results, 5 = Extreme opportunity for lack of

diligence to affect the results.

Experience and expertise of replication team. Just as experience and expertise may be necessary to obtain reproducible results, expertise and experience may be important for conducting effective replications. We focused on the senior member of the replication team and created an aggregate score of 7 characteristics: position (undergraduate to professor), highest degree (high school to PhD or equivalent), self-rated domain expertise, self-rated method expertise, total number of publications, total number of peer-reviewed empirical articles, and citation impact.

Position of senior member of replication team. Reproducibility may be enhanced by having more seasoned researchers guiding the research process. Replication teams reported the position of the senior member of the team from: 7 = Professor (or equivalent), 6 = Associate Professor (or equivalent), 5 = Assistant Professor (or equivalent), 4 = Post-doc, Research Scientist, or Private Sector Researcher, 3 = Ph.D. student, 2 = Master's student, 1 = Undergraduate student, or other.

Highest degree of replication team's senior member. Replication teams reported the highest degree obtained by the senior member of the team from 4 = PhD/equivalent, 3 = Master's/equivalent, 2 = some graduate school, 1 = Bachelor's/equivalent.

Replication team domain expertise. Reproducibility may be stronger if the replication team is led by a person with high domain expertise in the topic of study. Replication teams self-rated the domain expertise of the senior member of the project on the following scale: 1 = No expertise - No formal training or experience in the topic area, 2 = Slight expertise - Researchers exposed to the topic area (e.g., took a class), but without direct experience researching it, 3 = Some expertise - Researchers who have done research in the topic area, but have not published in it, 4 = Moderate expertise - Researchers who have previously published in the topic area of the selected effect, and do so irregularly, 5 = High expertise - Researchers who have previously published in the topic area of the selected effect, and do so regularly.

Replication team method expertise. Reproducibility may be stronger if the replication team is led by a person with high expertise in the methodology used for the study. Replication teams self-rated the domain expertise of the senior member of the project on the following scale: 1 = No expertise - No formal training or experience with the methodology, 2 = Slight expertise - Researchers exposed to the methodology, but without direct experience using it, 3 = Some expertise - Researchers who have used the methodology in their research, but have not published with it, 4 = Moderate expertise - Researchers who have previously published using the methodology of the selected effect, and use the methodology irregularly, 5 = High expertise - Researchers who have previously published using the methodology of the selected effect, and use the methodology regularly.

Replication team senior member's total publications and total number of peer-reviewed articles. All else being equal, more seasoned researchers may be better prepared to reproduce research results than more novice researchers. Replication teams self-reported the total number of publications and total number of peer-reviewed articles by the senior member of the team.

Institution prestige of replication 1st author and senior author. We followed the same

methodology for computing institution prestige for replication teams as we did for original author teams.

Citation impact of replication 1st author and senior author. Researchers who have conducted more research that has impacted other research via citation may have done so because of additional expertise and effectiveness in conducting reproducible research. Project members calculated the total citations of the 1st author and most senior member of the team via Google Scholar.

Self-assessed quality of replication. Lower quality replications may produce results less similar to original effects than higher quality replications. Replication teams are in the best position to know the quality of project execution, but are also likely to be ego invested in reporting high quality. Nonetheless, variation in self-assessed quality across teams may provide a useful indicator of quality. Also, some of our measures encouraged variation in quality reports by contrasting directly with the original study or studies in general. We created an aggregate score of four variables: self-assessed quality of implementation, self-assessed quality of data collection, self-assessed similarity to original, and self-assessed difficulty of implementation. Future research may assess additional quality indicators from the public disclosure of methods to complement this assessment.

Self-assessed implementation quality of replication. Sloppy replications may be less likely to reproduce original results because of error and inattention. Replication teams self-assessed the qualify of the replication study methodology and procedure design in comparison to the original research by answering the following prompt: “To what extent do you think that the replication study materials and procedure were designed and implemented effectively? Implementation of the replication materials and procedure…” Responses were provided on a scale from 1 = was of much higher quality than the original study, 2 = was of moderately higher quality than the original study, 3 = was of slightly higher quality than the original study, 4 = was about the same quality as the original study, 5 = was of slightly lower quality than the original study, 6 = was of moderately lower quality than the original study, 7 = was of much lower quality than the original study.

Self-assessed data collection quality of replication. Sloppy replications may be less likely to reproduce original results because of error and inattention. Replication teams self-assessed the quality of the replication study data collection in comparison to the average study by answering the following prompt: “To what extent do you think that the replication study data collection was completed effectively for studies of this type?” Responses were provided on a scale from 1 = Data collection quality was much better than the average study, 2 = Data collection quality was better than the average study, 3 = Data collection quality was slightly better than the average study, 4 = Data collection quality was about the same as the average study, 5 = Data collection quality was slightly worse than the average study, 6 = Data collection quality was worse than the average study, 7 = Data collection quality was much worse than the average study.

Self-assessed replication similarity to original. It can be difficult to reproduce the conditions and procedures of the original research for a variety of reasons. Studies that are more similar to the original research may be more reproducible than those that are more dissimilar. Replication teams self-evaluated the similarity of the replication with the original by

answering the following prompt: “Overall, how much did the replication methodology resemble the original study?” Responses were provided on a scale from 1 = Not at all similar, 2 = Slightly similar, 3 = Somewhat similar, 4 = Moderately similar, 5 = Very similar, 6 = Extremely similar, 7 = Essentially identical.

Self-assessed difficulty of implementation. Another indicator of adherence to the original protocol is the replication team’s self-assessment of how challenging it was to conduct the replication. Replication teams responded to the following prompt: “How challenging was it to implement the replication study methodology?” Responses were provided on a scale from 1 = Extremely challenging, 2 = Very challenging, 3 = Moderately challenging, 4 = Somewhat challenging, 5 = Slightly challenging, 6 = Not at all challenging.

Sampling Frame and Selection Biases

We constructed a sampling frame and selection process to minimize selection biases and maximize generalizability of the accumulated evidence. However, to maintain high quality, we had to balance these goals with having a selection process that was flexible enough to match replication projects with teams that had relevant interests and expertise. Initially, a total of 60 articles were made available from the sampling frame, starting with the first article published in the first 2008 issue of each of the three journals. Articles were matched with replication teams until the remaining articles were difficult to match. If there were still interested teams, then another 10 articles from one or more of the three journals were made available.

The most prevalent reasons for failure to match an article with a team were feasibility constraints for conducting the research - such as a difficult to obtain sample (e.g., identical twins), difficult to administer study design (e.g., 2-year longitudinal study), or requirements for specialized equipment or expertise (e.g., fMRI). Finally, after being claimed, some studies were not completed because the replication teams ran out of time or could not devote sufficient resources to completing the study. These present threats to generalizability of these findings. Of the 158 articles that were made eligible for claiming from the sampling frame, 98 were replicated and were included in the final dataset (62%) of 100 replications. Two articles had 2 replications each. By journal, replications were completed for 39 of 64 (61%) articles from PSCI, 31 of 55 (56%) articles from JPSP, and 28 of 39 (72%) articles from JEP:LMC.

Guide to the Information Commons

There is a substantial collection of materials comprising this project that is publicly accessible for review, critique, and reuse. The following list of links are a guide to the major components.

1. RPP OSF Project: The main repository for all project content is here (https://osf.io/ezcuj/)
2. RPP Information Commons: The project background and instructions for replication teams is in the wiki of the main OSF project (https://osf.io/ezcuj/wiki/home/)
3. RPP Researcher Guide: Protocol for replications teams to complete a replication (https://osf.io/ru689/)

4. Master Data File: Aggregate data across replication studies (https://osf.io/5wup8/)
5. Master Analysis Scripts: Script for reproducing analyses for each replication (https://osf.io/fkmwg/); script for reproducing Reproducibility Project: Psychology findings (https://osf.io/ki26g/)

Methods and Results for Individual Experiments

All reports, materials, and data for each replication are available publicly. In a few cases, research materials could not be made available because of copyright. In those cases, a note is available in that project's wiki explaining the lack of access and how to obtain the materials. The following table provides quick links to the projects (with data and materials), final reports, and the R script to reproduce the key finding for all replication experiments.

OSF project	Final report	R script to reproduce key finding
A Roelofs	https://osf.io/janu3/	https://osf.io/64pz8/
AL Alter, DM Oppenheimer	https://osf.io/jym7h/	https://osf.io/5axfe/
AL Morris, ML Still	https://osf.io/5f42t/	https://osf.io/qg9j7/
B Dessalegn, B Landau	https://osf.io/83n4z/	https://osf.io/qmupg/
B Eitam, RR Hassin, Y Schul	https://osf.io/x75fq/	https://osf.io/bvgyq/
B Liefooghe, P Barrouillet, A Vandierendonck, V Camos	https://osf.io/7ebqj/	https://osf.io/69b27/
B Monin, PJ Sawyer, MJ Marquez	https://osf.io/a4fmg/	https://osf.io/27gpt/
BC Storm, EL Bjork, RA Bjork	https://osf.io/byxjr/	https://osf.io/xsmzb/
BK Payne, MA Burkley, MB Stokes	https://osf.io/79y8g/	https://osf.io/u23g9/
C Farris, TA Treat, RJ Viken, RM McFall	https://osf.io/5u4km/	https://osf.io/ihcrs/
C Janiszewski, D Uy	https://osf.io/ehjdm/	https://osf.io/8qc4x/
C McKinstry, R Dale, MJ Spivey	https://osf.io/pu9nb/	https://osf.io/8hurj/
C Mitchell, S Nash, G Hall	https://osf.io/beckg/	https://osf.io/n539q/
CJ Berry, DR Shanks, RN Henson	https://osf.io/yc2fe/	https://osf.io/9ivaj/
CJ Soto, OP John, SD Gosling, J Potter	https://osf.io/6zdct/	https://osf.io/3y9sj/
CP Beaman, I Neath, AM Surprenant	https://osf.io/a6mje/	https://osf.io/pmhd7/
CR Cox, J Arndt, T Pyszczynski, J Greenberg, A Abdollahi, S Solomon	https://osf.io/uhnd2/	https://osf.io/fg2u9/
CS Dodson, J Darragh, A Williams	https://osf.io/b9dpu/	https://osf.io/dctav/

D Albarracín, IM Handley, K Noguchi, KC McCulloch, H Li, J Leeper, RD Brown, A Earl, WP Hart	https://osf.io/2pbaf/	https://osf.io/gtewj/
D Albarracín, IM Handley, K Noguchi, KC McCulloch, H Li, J Leeper, RD Brown, A Earl, WP Hart	https://osf.io/tarp4/	https://osf.io/256xy/
D Ganor-Stern, J Tzelgov	https://osf.io/7mgwh/	https://osf.io/txukv/
D Mirman, JS Magnuson	https://osf.io/r57hu/	https://osf.io/tjzqr/
DA Armor, C Massey, AM Sackett	https://osf.io/8u5v2/	https://osf.io/esa3j/
DB Centerbar, S Schnall, GL Clore, ED Garvin	https://osf.io/wcgx5/	https://osf.io/g29pw/
DM Amodio, PG Devine, E Harmon-Jones	https://osf.io/ysxmf/	https://osf.io/9gky5/
DR Addis, AT Wong, DL Schacter	https://osf.io/9ayxi/	https://osf.io/gfn65/
E Harmon-Jones, C Harmon-Jones, M Fearn, JD Sigelman, P Johnson	https://osf.io/zpwne/	https://osf.io/79ctv/
E Nurmsoo, P Bloom	https://osf.io/ictp5/	https://osf.io/ewtn6/
E van Dijk, GA van Kleef, W Steinel, I van Beest	https://osf.io/jyq3t/	https://osf.io/cxwev/
E Vul, H Pashler	https://osf.io/7kimb/	https://osf.io/8twa9/
E Vul, M Nieuwenstein, N Kanwisher	https://osf.io/jupew/	https://osf.io/2mcdv/
EJ Masicampo, RF Baumeister	https://osf.io/897ew/	https://osf.io/4tb8a/
EP Lemay, MS Clark	https://osf.io/efjn3/	https://osf.io/nhsdq/
EP Lemay, MS Clark	https://osf.io/mv3i7/	https://osf.io/wb4vd/
G Hajcak, D Foti	https://osf.io/83tsz/	https://osf.io/vjb2a/
G Tabibnia, AB Satpute, MD Lieberman	https://osf.io/56fmw/	https://osf.io/u5g9n/
GA Alvarez, A Oliva	https://osf.io/dm2kj/	https://osf.io/xgdqy/
GP Lau, AC Kay, SJ Spencer	https://osf.io/ndhwk	https://osf.io/cwkzu/
H Ersner-Hershfield, JA Mikels, SJ Sullivan, LL Carstensen	https://osf.io/4wskd/	https://osf.io/qedt9/
J Correll	https://osf.io/hzka3/	https://osf.io/476wy/
J Förster, N Liberman, S Kuschel	https://osf.io/sxnu6/	https://osf.io/h2r9c/
J Winawer, AC Huk, L Boroditsky	https://osf.io/ertbg/	https://osf.io/efu3h/
JA Richeson, S Trawalter	https://osf.io/phwi4/	https://osf.io/wi6hv/
JE Marsh, F Vachon, DM Jones	https://osf.io/sqcwk/	https://osf.io/pfmwj/

JI Campbell, ND Robert	https://osf.io/bux7k/	https://osf.io/z75yu/
JJ Exline, RF Baumeister, AL Zell, AJ Kraft, CV Witvliet	https://osf.io/es7ub/	https://osf.io/jfigk/
JL Risen, T Gilovich	https://osf.io/wvcgb/	https://osf.io/itc9q/
JL Tracy, RW Robins	https://osf.io/9uqxr/	https://osf.io/k7huw/
JR Crosby, B Monin, D Richardson	https://osf.io/nkaw4/	https://osf.io/3nay6/
JR Schmidt, D Besner	https://osf.io/bskwq/	https://osf.io/ktgnq/
JS Nairne, JN Pandeirada, SR Thompson	https://osf.io/v4d2b/	https://osf.io/witg3/
JT Larsen, AR McKibban	https://osf.io/h4cbg/	https://osf.io/df7cj/
K Fiedler	https://osf.io/vtz2i/	https://osf.io/4m8ir/
K Oberauer	https://osf.io/n32zj/	https://osf.io/vhzi6/
KA Ranganath, BA Nosek	https://osf.io/9xt25/	https://osf.io/m4xp8/
KD Vohs, JW Schooler	https://osf.io/5bn6g/	https://osf.io/eyk8w/
KE Stanovich, RF West	https://osf.io/p3gz2/	https://osf.io/jv4tw/
KL Blankenship, DT Wegener	https://osf.io/v3e2z/	https://osf.io/4vuhw/
KR Morrison, DT Miller	https://osf.io/2jwi6/	https://osf.io/hau4p/
L Demany, W Trost, M Serman, C Semal	https://osf.io/wx74s/	https://osf.io/cfbk8/
L Sahakyan, PF Delaney, ER Waldum	https://osf.io/kcwfa/	https://osf.io/2hasj/
LE Williams, JA Bargh	https://osf.io/7uh8g/	https://osf.io/85bnh/
LS Colzato, MT Bajo, W van den Wildenberg, D Paolieri, S Nieuwenhuis, W La Heij, B Hommel	https://osf.io/a5ukz/	https://osf.io/kb59n/
M Bassok, SF Pedigo, AT Oskarsson	https://osf.io/irgbs/	https://osf.io/25vhj/
M Couture, D Lafond, S Tremblay	https://osf.io/qm5n6/	https://osf.io/3zg7e/
M Koo, A Fishbach	https://osf.io/68m2c/	https://osf.io/p5i9j/
M Reynolds, D Besner	https://osf.io/fkcn5/	https://osf.io/yscmg/
M Tamir, C Mitchell, JJ Gross	https://osf.io/7i2tf/	https://osf.io/mwgub/
MD Henderson, Y de Liver, PM Gollwitzer	https://osf.io/cjr7d/	https://osf.io/b2ejv/
MJ Yap, DA Balota, CS Tse, D Besner	https://osf.io/dh4jx/	https://osf.io/nuab4/
N Epley, S Akalis, A Waytz, JT Cacioppo	https://osf.io/m5a2c/	https://osf.io/utcr3/
N Halevy, G Bornstein, L Sagiv	https://osf.io/sjwcd/	https://osf.io/7xyi5/
N Janssen, FX Alario, A Caramazza	https://osf.io/e3ry5/	https://osf.io/7cab3/

N Janssen, W Schirm, BZ Mahon, A Caramazza	https://osf.io/ka5vp/	https://osf.io/iwaqf/
N Shnabel, A Nadler	https://osf.io/hxbvn/	https://osf.io/5bwva/
NB Turk-Browne, PJ Isola, BJ Scholl, TA Treat	https://osf.io/ktnmc/	https://osf.io/gpvrm/
NO Rule, N Ambady	https://osf.io/4peq6/	https://osf.io/2bu9s/
P Bressan, D Stranieri	https://osf.io/7vriw/	https://osf.io/2a5ru/
P Bressan, D Stranieri	https://osf.io/7vriw/	https://osf.io/47cs8/
P Fischer, S Schulz-Hardt, D Frey	https://osf.io/5afur/	https://osf.io/bajxq/
P Fischer, T Greitemeyer, D Frey	https://osf.io/9pnct/	https://osf.io/7htc9/
PA Goff, CM Steele, PG Davies	https://osf.io/7q5us/	https://osf.io/xfj5w/
PA White	https://osf.io/x7c9i/	https://osf.io/ygh35/
PW Eastwick, EJ Finkel	https://osf.io/5pjsn/	https://osf.io/x3hbe/
S Farrell	https://osf.io/tqf2u/	https://osf.io/nmpdc/
S Forti, GW Humphreys	https://osf.io/nhqgs/	https://osf.io/jknef/
S Pacton, P Perruchet	https://osf.io/asn7w/	https://osf.io/3kn4c/
S Schnall, J Benton, S Harvey	https://osf.io/2dem3/	https://osf.io/pkaqw/
SE Palmer, T Ghose	https://osf.io/4ynbx/	https://osf.io/jnqky/
SJ Heine, EE Buchtel, A Norenzayan	https://osf.io/g4hn3/	https://osf.io/akv6y/
SK Moeller, MD Robinson, DL Zabelina	https://osf.io/7dybc/	https://osf.io/uevha/
SL Murray, JL Derrick, S Leder, JG Holmes	https://osf.io/3hndq/	https://osf.io/9ue7j/
SM McCrea	https://osf.io/ytxgr/	https://osf.io/7pdh8/
T Goschke, G Dreisbach	https://osf.io/pnius/	https://osf.io/mvdsw/
T Makovski, R Sussman, YV Jiang	https://osf.io/xtcuv/	https://osf.io/saq6x/
TJ Pleskac	https://osf.io/gyn9e/	https://osf.io/scqrd/
V LoBue, JS DeLoache	https://osf.io/5ygej/	https://osf.io/p67kr/
V Purdie-Vaughns, CM Steele, PG Davies, R Ditlmann, JR Crosby	https://osf.io/3rxvs/	https://osf.io/5i8tu/
X Dai, K Wertenbroch, CM Brendl	https://osf.io/js7gd/	https://osf.io/aigrv/
Z Estes, M Verges, LW Barsalou	https://osf.io/b7zek/	https://osf.io/4di3e/

Statistical Analyses

The reproducibility of psychological science was evaluated using significance and *p*-values, effect sizes, subjective assessments of replication teams, and meta-analysis of original effect sizes, replication effect sizes, and the difference of effect size between original and replication study.

Significance and *p*-values. Assuming a two-tailed test and significance or alpha level of .05, all test results of original and replication studies were classified as statistically significant (*p*-value ≤ 0.05) and non-significant (*p* > .05). However, original studies that interpreted non-significant *p*-values as significant were coded as significant (4 cases, all with *p*-values < .06). Using the non-significant *p*-values of the replication studies only, using Fisher's (*38*) method we tested the hypothesis that these studies had 'no evidential value' (i.e., the null-hypothesis of zero-effect holds for all these studies). The hypothesis that the proportions of statistically significant results are equal was tested using the McNemar test for paired nominal data, and a confidence interval of the reproducibility parameter was calculated. Second, we compared the central tendency of the distribution of *p*-values of original and replication studies using the Wilcoxon signed-rank test and the *t*-test for dependent samples. For both tests we only used complete data (i.e., study-pairs for which both *p*-values were available).

Effect sizes. We transformed all effect sizes into correlation coefficients whenever possible. Correlation coefficients have several advantages over other effect size measures, such as Cohen's *d*. Correlation coefficients are bounded, well-known, and therefore more readily interpretable. Most importantly for our purposes, analysis of correlation coefficients is straightforward because, after applying the Fisher transformation, their standard error is only a function of sample size. Formulas and code for converting test statistics *z*, *F*, *t*, and χ^2 into correlation coefficients are provided in [A3]. To be able to compare and analyze correlations across study-pairs, the original study's effect size was coded as positive; the replication study's effect size was coded as negative only if the replication study's effect was opposite to that of the original study.

Effect sizes were compared using four tests. The central tendency of the effect size distributions of original and replication studies were compared using both a paired two-sample *t*-test and the Wilcoxon signed-rank test. Third, we computed the proportion of study-pairs in which the effect of the original study was stronger than in the replication study, and tested the hypothesis that this proportion is .5. For this test we used the data for which effect size measures were available but no correlation coefficient could be computed (e.g., if a regression coefficient was reported, but not its test statistics). Fourth, we calculated 'coverage', or the proportion of study-pairs in which the effect of the original study was in the confidence interval of the effect of the replication study, and compared this with the expected proportion using a goodness-of-fit χ^2-test. We carried out this test on the subset, further called MA, of study-pairs where both the correlation coefficient and its standard error could be computed. Standard errors could only be computed if test statistics were *r*, *t*, or $F(1,df_2)$. The

expected proportion is the sum over expected probabilities across study-pairs. The test assumes the same population effect size for original and replication study in the same study-pair (see [A4] for computational details on the test). For those studies that tested the effect with $F(df_1 > 1, df_2)$ or χ^2, we verified coverage using other statistical procedures (see [A5]).

Meta-analysis combining original and replication effects. Fixed-effect meta-analyses were conducted in metafor on Fisher-transformed correlations for all study-pairs in subset MA, and on study-pairs with the odds ratio as the dependent variable. The number of times the CI of all these meta-analyses contained 0 was calculated, whereas only estimated effect sizes of those in subset MA were averaged and analyzed by discipline.

Subjective assessment of "Did it replicate?" In addition to the quantitative assessments of replication and effect estimation, we conducted a subjective assessment of whether the replication provided evidence of replicating the original result. In some cases, the quantitative data anticipates a straightforward subjective assessment of replication. But for more complex designs, such as multivariate interaction effects, the quantitative analysis may not provide a simple interpretation. For subjective assessment, replication teams answered "yes" or "no" to the question "Did your results replicate the original effect?". Also, replication teams answered two questions "How much did the key effect in the replication resemble the key effect in the original study?" and "Overall, how much did the findings in the replication resemble the findings in the original study?" The second question assessed all findings the replication team reported, allowing for a measure that encompassed more than just the key effect.

Meta-analysis of all original study effects, and of all replication study effects. Two random-effects meta-analyses were run using REML estimation in metafor, one on effect sizes of original and one on effect sizes of replication studies, both of studies in set MA. We ran three models; one without any predictor, one with studies' standard error as predictor, and one with standard error and discipline as predictor. Standard error was added to examine small-study effects. A positive effect of standard error on effect size indicates that studies' effect sizes are positively associated with their sample sizes. The results of this one-tailed test, also known as Egger's test, is often used as test of publication bias. Discipline is a categorical variable with categories JPSP-social (= reference category), JEP:LMC-cognitive, PSCI-social, PSCI-cognitive, and PSCI-other.

Meta-analysis of difference of effect size between original and replication study. The dependent variable was the difference of Fisher-transformed correlations (original – replication), with variance equal to the sum of variances of the correlation of the original and of the replication study. Several random-effect meta-analyses were run using REML estimation in metaphor. First, the intercept-only model was estimated; the intercept denotes the average difference effect size between original and replication study. Second, to test for small study effects, we added the standard error of the original study as a predictor, akin to

Egger's test; a positive effect is often interpreted as evidence for publication bias. Our third model tested the effect of discipline.

Analysis of moderators. We correlated replication success ($p < .05$) Fisher-transformed difference in effect size between original and replication studies estimated meta-analysis across study-pairs, whether the original effect size was in the replication 95% CI, and subjective assessment of replication success with six indicators of the original study (original *p*-value, original effect size, original sample size, importance of the effect, surprising effect, experience and expertise of original team) and seven indicators of the replication study (replication *p*-value, replication effect size, replication power (based on original effect size), replication sample size, challenge of conducting replication, experience and expertise of replication team, self-assessed quality of replication; see Table 2). As follow-up, we did the same with the individual indicators comprising the moderator variables (Tables S4 and S5). Those are detailed in the methods section above.

Results

Preliminary analyses

The input of our analyses were the *p*-values (DH and DT in the Master Data File), their significance (columns EA and EB), effect sizes of both original and replication study (columns DJ and DV), which effect size was larger (column EC), direction of the test (column BU), and whether the sign of both studies' effects was the same or opposite (column BT). First, we checked the consistency of *p*-value and test statistics whenever possible (i.e., when all were provided), by recalculating the *p*-value using the test statistics. We used the recalculated *p*-values in our analysis, with a few exceptions (see [A1] for details on the recalculation of *p*-values). These *p*-values were used to code the statistical (non-)significance of the effect, with the exception of four effects with *p*-values slightly larger than .05 (.0503, .0509, .0514, and .0516) that were interpreted as significant; these studies were treated as significant. We ended up with 99 study-pairs with complete data on *p*-values, and 100 study-pairs with complete data on the significance of the replication effect.

Table S1. Statistical results (statistically significant or not) of original and replication studies.

Results

		Replication	
		Nonsignificant	Significant
Original	Nonsignificant	2	1
	Significant	62	35

The effect sizes ("correlation per df") were computed using the test statistics (see [A3] for details on the computation of effect sizes), taking the sign of observed effects into account. Because effect size could not be computed for three study-pairs, we ended up with 97

study-pairs with complete data on effect size. Of the three missing effect sizes, for two could be determined which effect size was larger, hence we ended up with 99 study-pairs with complete data on the comparison of the effect size. Depending on the assessment of replicability, different study-pairs could be included. Seventy-four study-pairs could be included in subset MA, 76 (74+2) could be used to test if the study-pair's meta-analytic estimate was larger than zero, and 95 (76+19) could be used to determine if the CI of the replication contained the effect size of the original study (see end of [A3] for an explanation).

Evaluating replication effect against null hypothesis of no effect.

See [A2] for details. Table S1 shows the statistical significance of original and replication studies. Of the original studies, 97% were statistically significant, as opposed to 36.0% (CI = [26.6%, 46.2%]) of replication studies, which corresponds to a significant change (McNemar test, $\chi^2(1) = 59.1$, $p < .001$).

Proportions of statistical significance of original and replication studies for the three journals JPSP, JEP, PSCI were .969 and .219, .964 and .464, .975 and .4, respectively. Of 97 significant original studies, 36.1% were statistically significant in the replication study. The hypothesis that all 64 statistically non-significant replication studies came from a population of true negatives can be rejected at significance level .05, but not at .01 ($\chi^2(128) = 155.83$, $p = 0.048$).

The density and cumulative *p*-value distributions of original and replication studies are presented in Figures S1 and S2 respectively. The means of the two *p*-value distributions (.028 and .302) were different from each other ($t(98) = -8.22$, $p < .001$; W = 2406, $p < .001$). Quantiles are .00042, .0069, .023 for the original, and .0078, .20, .54 for the replication studies.

Figure S1: Cumulative *p*-value distributions of original and replication studies.

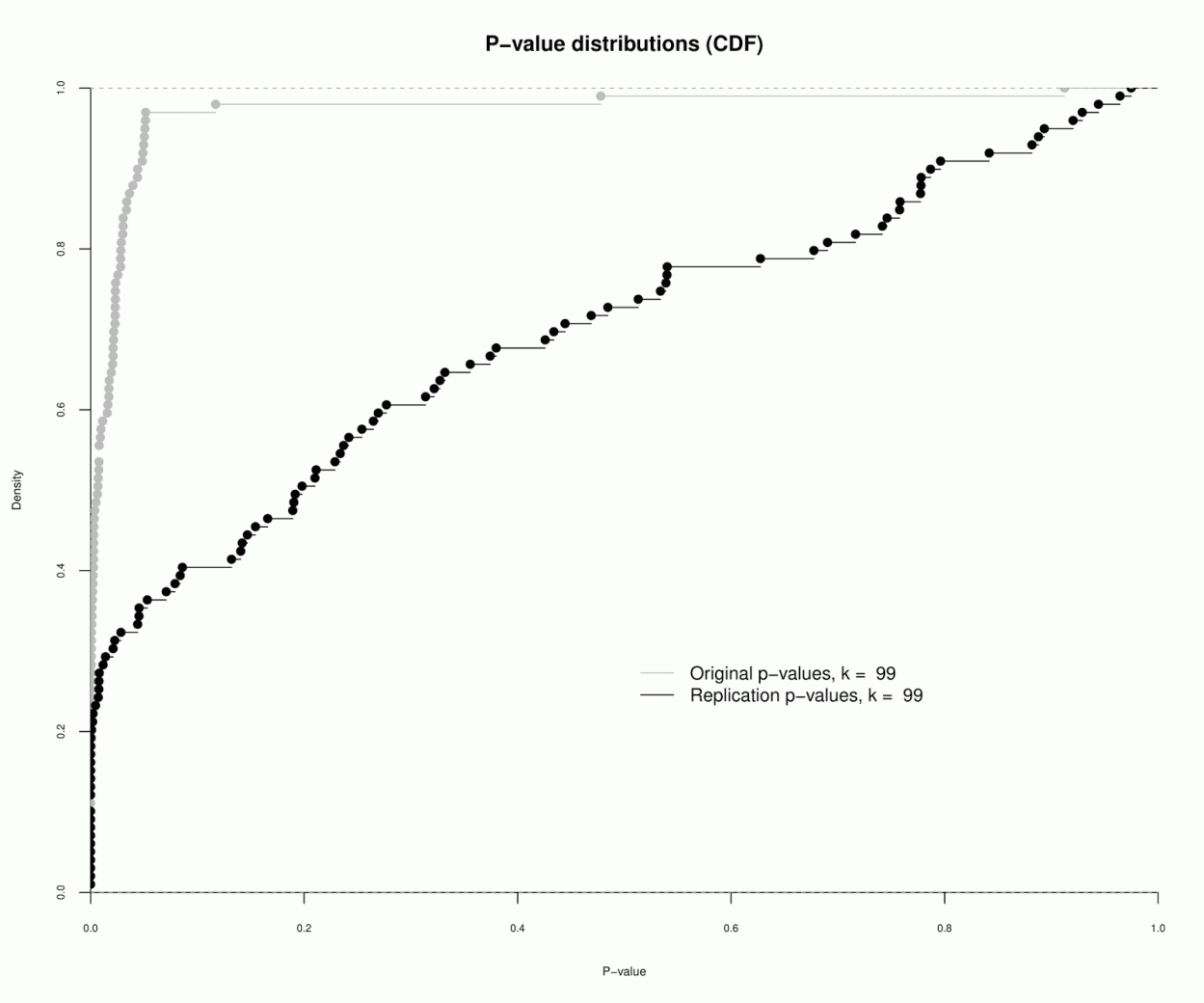
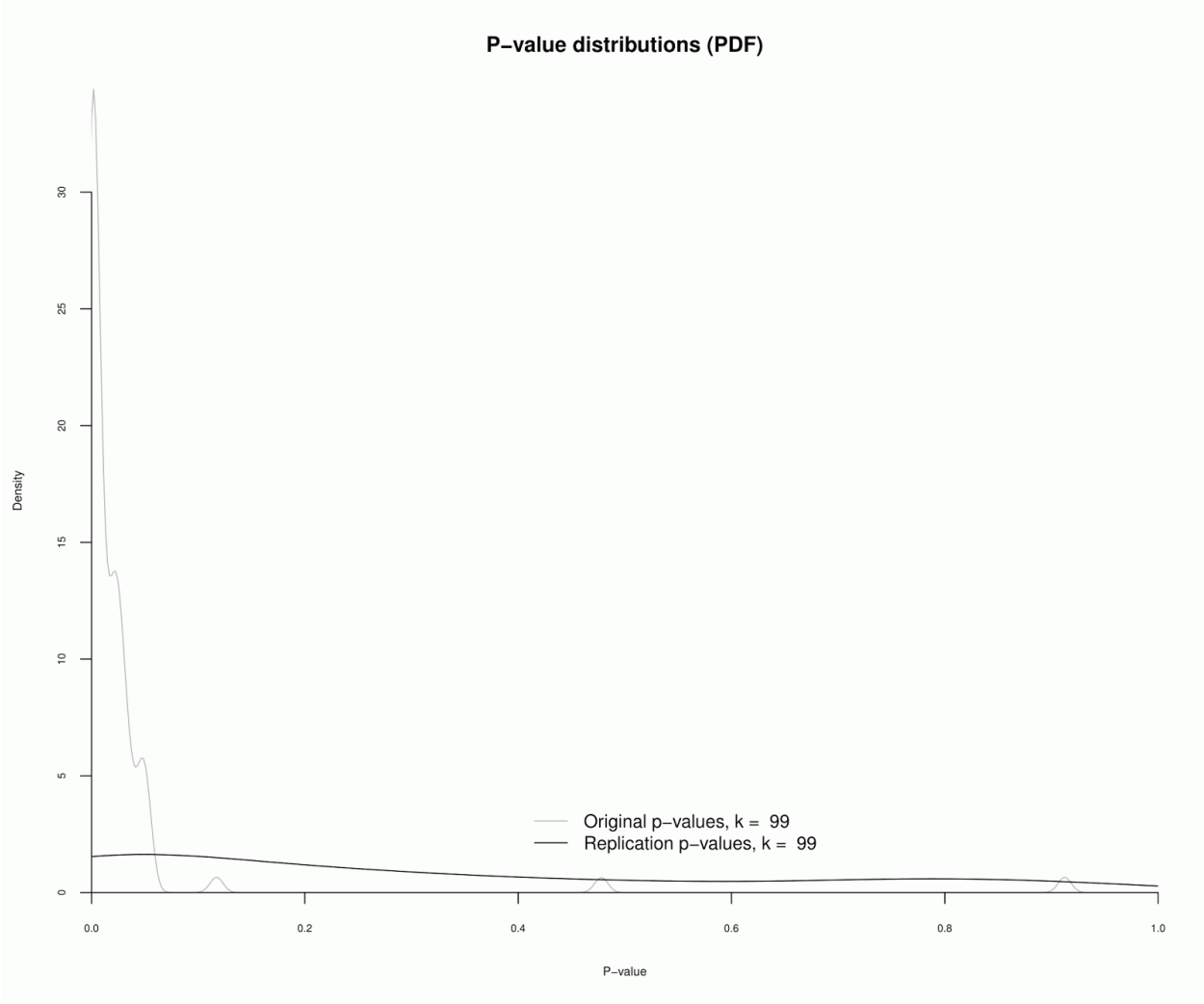


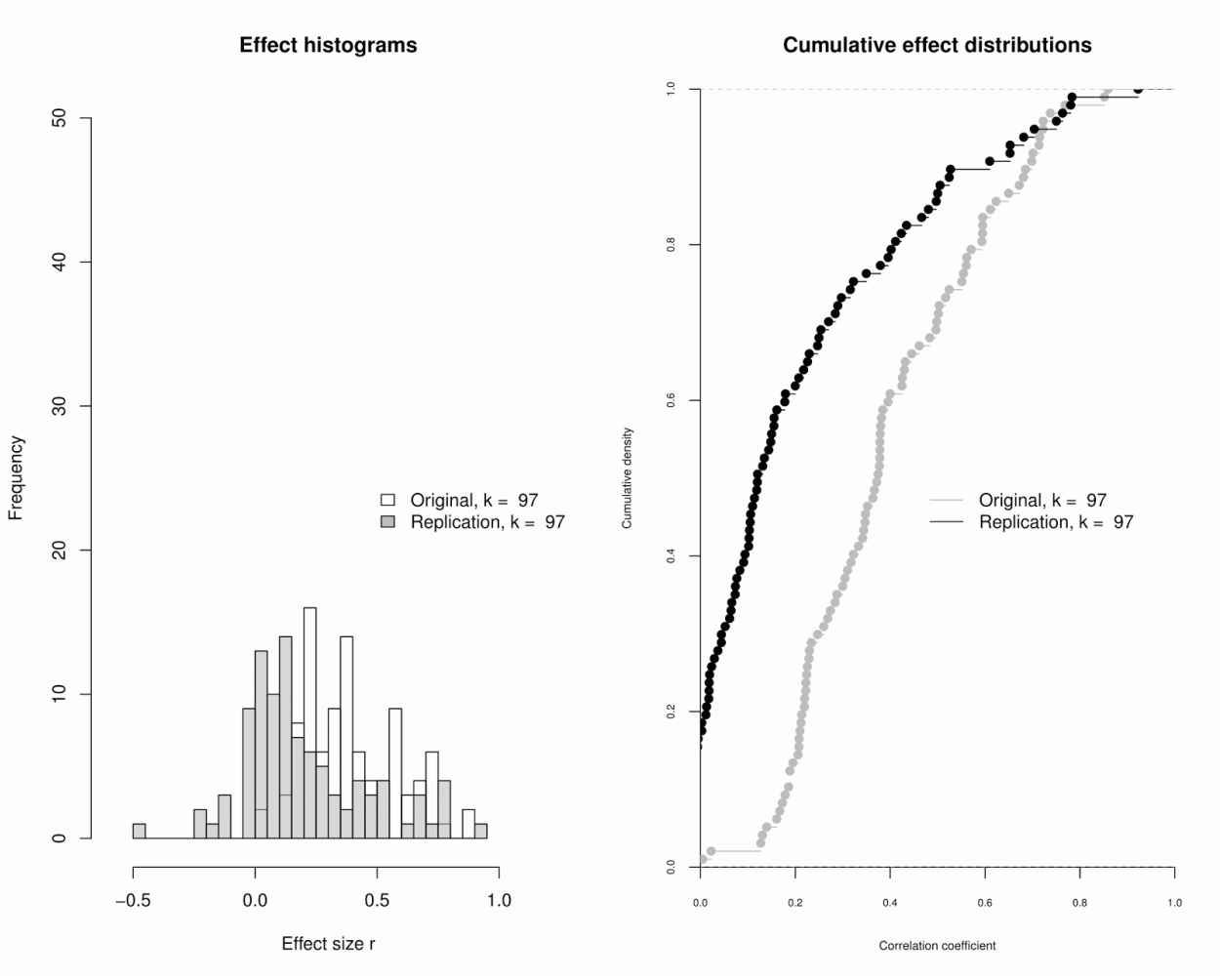
Figure S2: Density *p*-value distributions of original and replication studies



Comparing original and replication effect sizes.

See [A3] and [A6] for details. For 97 study pairs effect size correlations could be computed. Figure S3 (left) shows the distribution of effect sizes of original and replication studies, and the corresponding cumulative distribution functions (right). The mean effect sizes of both distributions (M = .396 [SD = .193]; M = .198 [SD = .255]) were different from each other ($t(96) = 9.33$, $p < .001$; $W = 7132$, $p < .001$). Of those 99 studies that reported an(y) effect size in both original and replication study, 82 reported a stronger effect size in the replication study (82.8%; $p < .001$, binomial test). Original and replication effect sizes were positively correlated (Spearman's $r = .51$, $p < .001$).

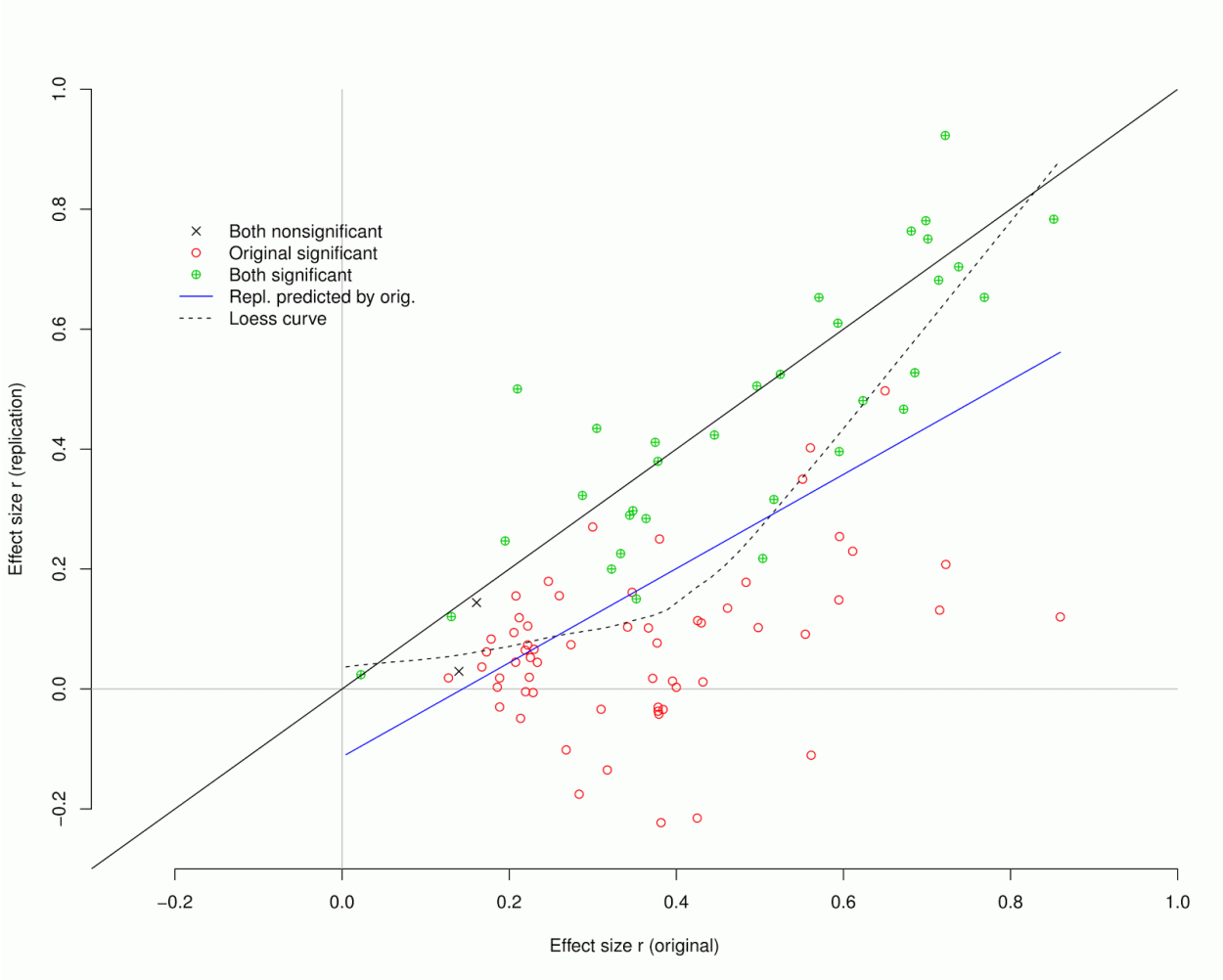
Figure S3: Distributions (left) and cumulative distribution functions of effect sizes of original and replication studies.



Evaluating replication effect against original effect size.

For the subset of 74 studies where the standard error of the correlation could be computed, it was expected that 78.5% of CIs of the replication study contained the effect size of the original study; however, only 41.9% (31 out of 74) of CIs contained the original effect size ($p < .001$) (see [A4] for details). For the subset of 17 and 4 studies with test statistics $F(df_1 > 1, df_2)$ and χ^2, respectively, 66.7% of the confidence intervals contained the effect size of the original study (see [A5] for details). This results in an overall success rate of 47.4%. Figure S4 depicts effect sizes of study-pairs for which correlations could be calculated, and codes significance of effect sizes as well.

Figure S4: Correlations of both original and replication study, coded by statistical significance. Identical values are indicated by the black diagonal line, whereas the blue and dotted line show the replication correlations as predicted by a linear model and loess, respectively.



Combining original and replication effect sizes for cumulative evidence.

See [A7] for details. For 74 study-pairs a meta-analysis could be conducted on the Fisher-transformed correlation scale. In 52 out of 74 pairs the null-hypothesis of no effect was rejected (70.3%). The average correlation, after transforming back the Fisher-transformed estimate, was .317 (*SD* = .228). However, the results differed across discipline; average effect size was smaller for JPSP (*M* = .142, *SD* = .084) than for the other four disciplines, and the percentage of meta-analytic effectsrejecting the null-hypothesis was also lowest for JPSP (42.8%; see Table 1). As noted in the main text, the interpretability of these meta-analytic estimates is qualified by the possibility of publication bias inflating the original effect sizes.

Subjective assessment of "Did it replicate?"

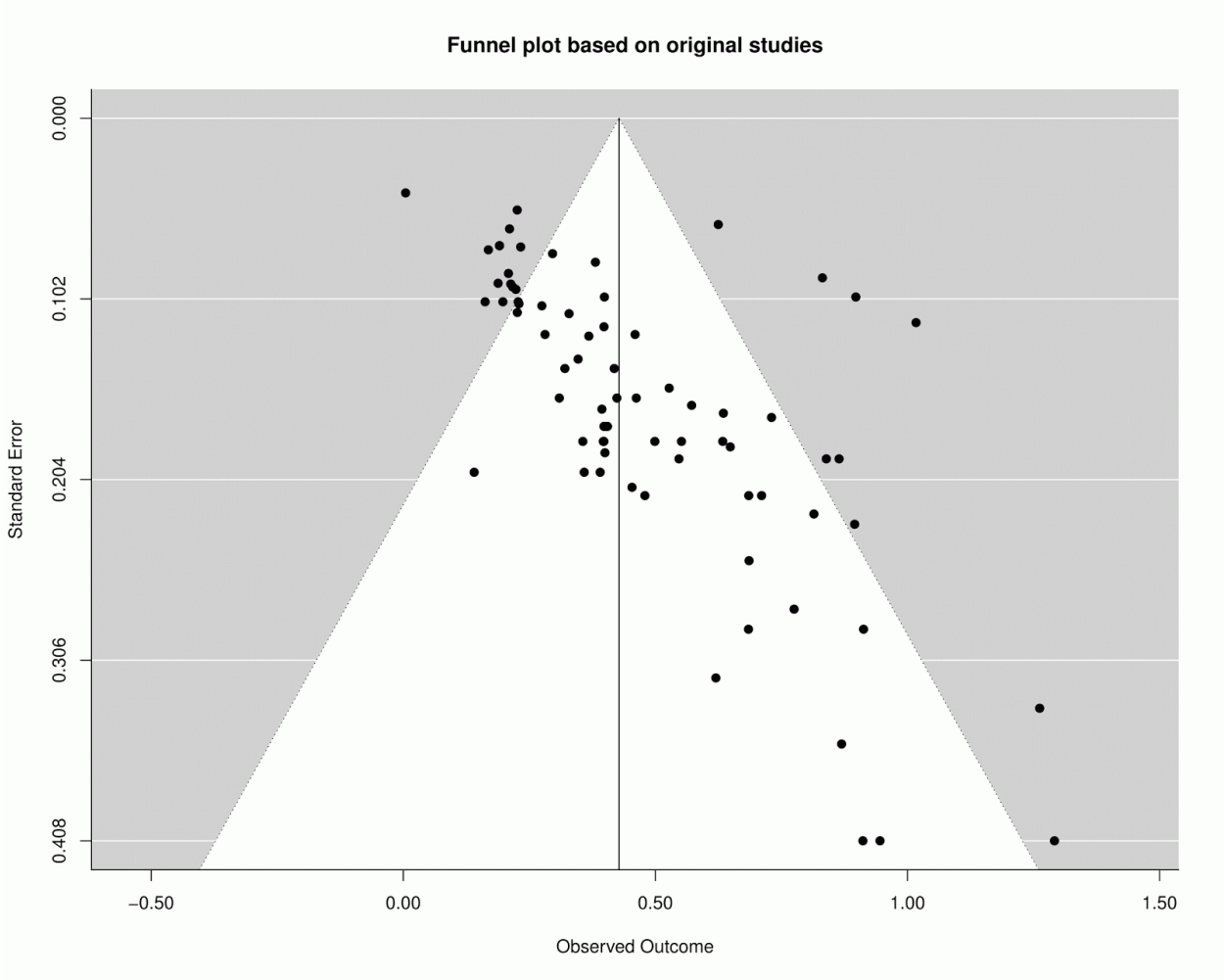
Replication teams provided a dichotomous yes/no assessment of whether the effect replicated or not (Column BX). Assessments were very similar to evaluations by significance testing ($p < .05$) including two original null results being interpreted as successful replications when the replication was likewise null, and one original null results being interpreted as a failed replication when the replication showed a significant effect. One positive, original effect that did not meet the $p < .05$ criterion was subjectively assessed as a successful replication resulting in 38 assessments of successful replication (38 of 100; 38%).

There are three subjective variables assessing replication success. Additional analyses can be conducted on replication teams' assessments of the extent to which key effect and overall findings resemble the original results (Columns CR and CQ).

Meta-analysis of all original study effects, and of all replication study effects.

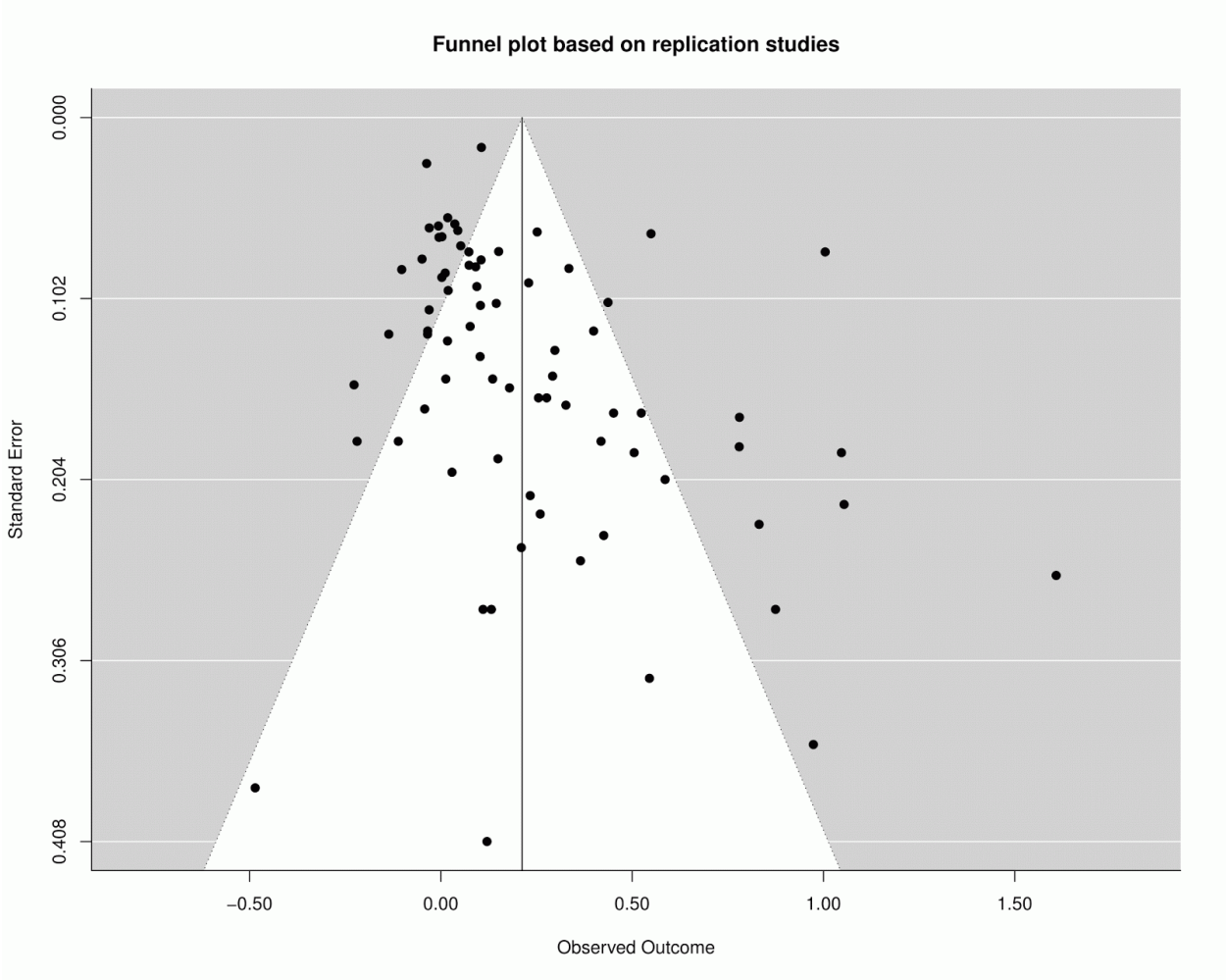
See [A7] for details. The meta-analysis on all original study effect sizes showed significant ($Q(73) = 310.45$, $p < .001$) and large heterogeneity ($\hat{\tau} = .19$, $I^2 = 73.8\%$), with average effect size equal to .43 ($z = 14.79$, $p < .001$). The average effect size differed across disciplines ($Q_M(4) = 14.08$, $p = .0070$), with effect size in JPSP (.29) being significantly smaller than in JEP (.52; $z = 3.13$, $p = .0017$). PSCI-Cog (.58; $z = 2.59$, $p = .0097$), PSCI-oth (.54; $z = 2.11$, $p = .035$), but not PSCI-Soc (.42; $z = 1.72$, $p = .086$). The effect of the original studies' standard error on effect size was large and highly significant ($b = 2.33$, $z = 6.07$, $p < .001$). Figure S5 shows the funnel plot of the meta-analysis without predictors. After controlling for study's standard error, there was no longer an effect of discipline on effect size ($\chi^2(4) = 5.29$, $p = .26$).

Figure S5: Funnel plot of the meta-analysis on the original study's effect size.



The same meta-analysis on replication studies' effect sizes showed significant ($Q(74)$ = 468.23, $p < .001$) and large heterogeneity ($\hat{\tau}$ =.27, I^2 = 90.7%), with average effect size equal to .21 (z = 5.97, $p < .001$). The average effect size again differed across disciplines ($Q_M(4)$ = 15.24, p = .0042). Average effect size in JPSP did not differ from 0 (.042; z = .71, p = .48), and was significantly smaller than average effect size in JEP (.28; z = 2.80, p = .0051), PSCI-Cog (.45; z = 3.29, p = .0010), and PSCI-Soc (.26; z = 2.52, p = .0119), but was not significantly smaller than effect size in PSCI-oth (.14; z = 0.60, p = .55). The effect of the standard error of the replication study was large and highly significant (b = 1.74, z = 3.74, $p < .001$). Figure S6 shows the corresponding funnel plot. The effect of discipline remained after controlling for the standard error of the replication study ($\chi^2(4)$ = 13.01, p = .011).

Figure S6: Funnel plot of the meta-analysis on the replication study's effect size.

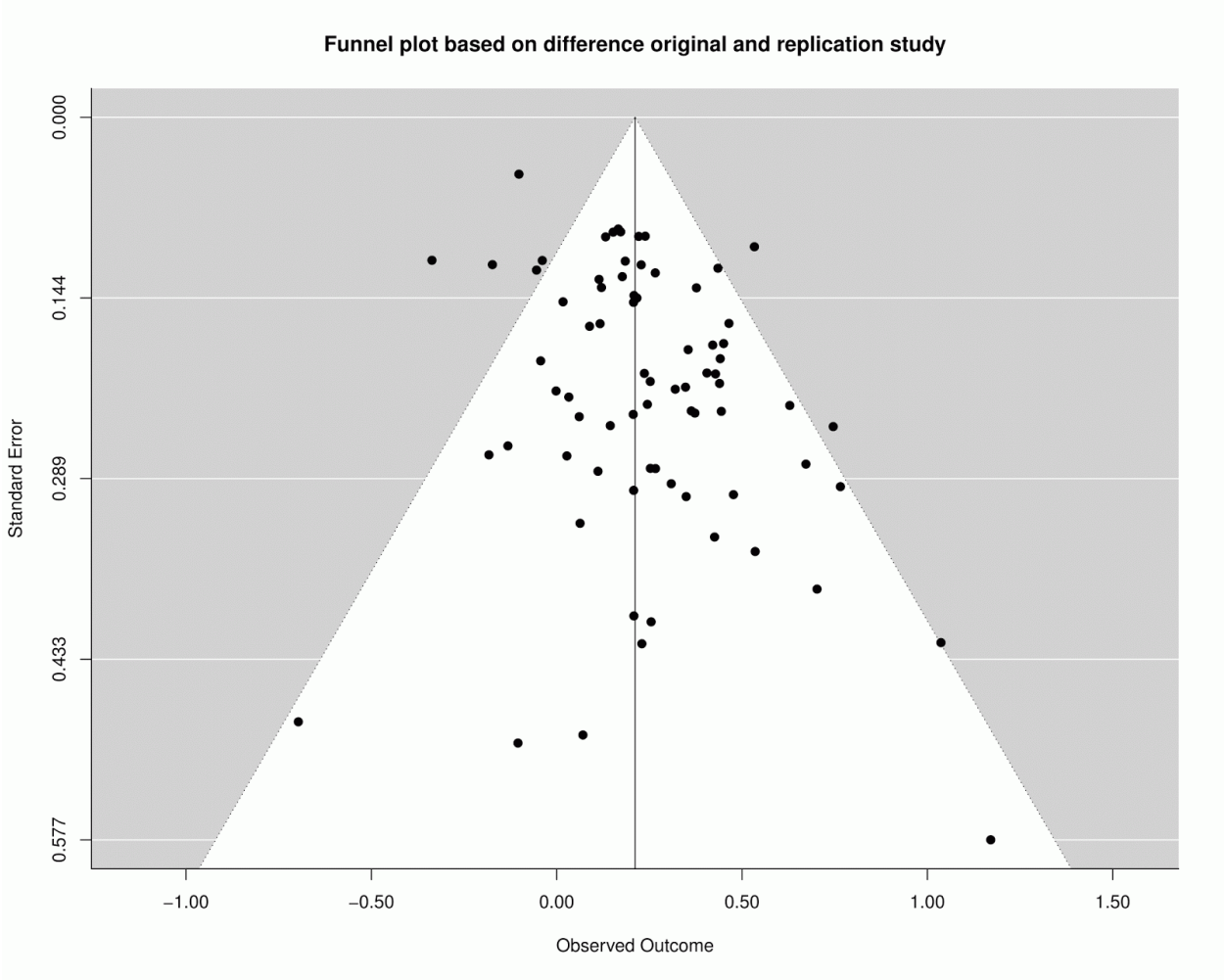


Meta-analysis of difference of effect size between original and replication study

The null-model without predictors yielded an average estimated difference in effect size equal to .21 (z = 7.51, p < .001) in favor of the original study. The null-hypothesis of homogenous difference in effect sizes was rejected ($Q(73)$ = 151.46, p < .001), with medium observed heterogeneity ($\hat{\tau}$ =.148, I^2 = 47.2%). Via Egger's test, precision of the original study was associated with the difference in effect size (b = .82, z = 1.87, one-tailed p = .031), hence imprecise original studies (large standard error) yielded larger differences in effect size between original and replication study. This is confirmed by the funnel plot in Figure S7. Discipline was not associated with the difference in effect size, $\chi^2(4)$ = 5.91, p = .21, (i.e., the average difference in effect size was equal for JPSP, JEP, PSCI-soc, PSCI-cog, and PSCI-other). Also, after controlling for the effect of the standard error of the original study, no differences between disciplines were observed ($\chi^2(4)$ = 6.04, p = .20). No moderating effects

were observed for: importance of the effect (b = -.012, p = .73), surprising effect (b = .0003, p = .99), experience and expertise of original team (b = -.0014, p = .97), challenge of conducting replication (b = 0.023, p = .50), and self-assessed quality of replication (b = -.039, p = .49). However, a positive effect of experience and expertise of replication team was observed (b = .13, p = .013), meaning that the difference in effect size was *higher* for replication teams with more experience and expertise.

Figure S7: Funnel plot of meta-analysis on difference in effect size (original – replication).



Moderator Analyses

The main text reports correlations between five reproducibility indicators and aggregate variables of original and replication study characteristics. Below are correlations among the five reproducibility indicators (Table S3), correlations of individual characteristics of original studies with reproducibility indicators (Table S4), and correlations of individual characteristics of replication studies with reproducibility indicators (Table S5).

Table S3. Spearman's rank order correlations among reproducibility indicators

	Replications p < .05 in original direction	Effect Size Difference	Meta-analytic Estimate	original effect size within replication 95% CI	subjective "yes" to "Did it replicate?"
Replications p < .05 in original direction	.				
Effect Size Difference	-0.598	.			
Meta-analytic Estimate	0.606	-0.206	.		
original effect size within replication 95% CI	0.551	-0.484	0.525	.	
subjective "yes" to "Did it replicate?"	0.978	-0.571	0.558	0.587	.

Notes: Effect size difference (original - replication) computed after converting r's to Fischer's z. Notes: Four original results had p-values slightly higher than .05, but were considered positive results in the original article and are treated that way here. Exclusions (see SI [A3] for explanation): "replications p < .05" (3 excluded; n = 97), "effect size difference" (3 excluded; n = 97); "meta-analytic mean estimates" (26 excluded; n = 74); and, "% original effect size within replication 95% CI" (5 excluded, n=95).

Table S4. Spearman's rank-order correlations of reproducibility indicators with individual original study characteristics

	M	SD	Median	Range	Replications p < .05 in original direction	Effect Size Difference	Meta-analytic Estimate	original effect size within replication 95% CI	subjective "yes" to "Did it replicate?"
Original effect size	0.3942	0.2158	0.3733	.0046 to .8596	0.304	0.310	0.800	0.137	0.260
Original p-value	0.0283	0.1309	0.0069	0 to .912	-0.327	-0.074	-0.476	0.019	-0.294
Original df/N	2408	22994	55	7 to 230025	-0.187	-0.114	-0.479	-0.239	-0.183
Institution prestige of 1st author	3.78	1.49	3.45	1.28 to 6.74	-0.026	0.030	-0.041	-0.120	-0.037
Institution prestige of senior author	3.96	1.55	3.65	1.28 to 6.74	-0.049	-0.060	0.023	-0.083	-0.046
Citation impact of 1st author	3074	5341	1539	54 to 44032	0.117	-0.111	0.086	0.002	0.136
Citation impact of senior author	13656	17220	8475	240 to 86172	-0.093	-0.061	-0.156	-0.047	-0.089
Article citation impact	84.91	72.95	56	6 to 341	-0.013	-0.044	-0.126	-0.061	0.000
Internal conceptual replications	0.91	1.21	0	0 to 5	-0.164	0.031	-0.204	-0.067	-0.174
Internal direct replications	0.06	0.32	0	0 to 3	0.061	0.021	0.064	0.113	0.051
Perceived expertise required	2.25	1.2	2	1 to 5	-0.114	0.061	-0.016	-0.057	-0.044
Perceived opportunity for expectancy bias	1.74	0.8	2	1 to 4	-0.214	0.098	-0.361	-0.120	-0.184
Perceived opportunity for impact of lack of diligence	2.21	1.02	2	1 to 5	-0.194	0.088	-0.286	-0.024	-0.148
Surprising result	2.96	0.85	3	1.25 to 5.2	-0.258	0.061	-0.232	-0.068	-0.212
Exciting/important result	3.25	0.66	3.2	1.33 to 4.75	-0.143	0.043	-0.215	-0.021	-0.109

Table S5. Spearman's rank-order correlations of reproducibility indicators with individual replication study characteristics

	M	SD	Median	Range	Replications p < .05 in original direction	Effect Size Difference	Meta-analytic Estimate	original effect size within replication 95% CI	subjective "yes" to "Did it replicate?"
Institution prestige of 1st author	3.03	1.42	2.53	1.31 to 6.74	-0.220	0.120	-0.406	-0.253	-0.217
Institution prestige of senior author	3.02	1.39	2.61	1.31 to 6.74	-0.227	0.100	-0.394	-0.291	-0.223
Citation count of 1st author	570	1280	91	0 to 6853	0.064	-0.087	-0.018	0.233	0.030
Citation count of senior author	1448	2571	378	0 to 15770	-0.092	0.107	-0.022	0.052	-0.098
Position of senior member of replication team	2.9	1.89	2	1 to 7	-0.154	0.074	-0.233	-0.195	-0.131
Highest degree of senior member	1.23	0.62	1	1 to 4	-0.013	-0.033	-0.054	-0.159	0.005
Senior member's total publications	45.26	69.21	18.5	0 to 400	-0.033	0.080	0.070	0.063	-0.014
Domain expertise	3.17	1.09	3	1 to 5	0.076	0.035	0.150	0.176	0.117
Method expertise	3.44	1.09	3	1 to 5	-0.063	0.148	0.256	0.034	-0.063
Implementation quality	3.85	0.86	4	1 to 6	-0.058	0.083	-0.097	0.044	-0.026
Data collection quality	3.60	1.00	4	1 to 6	-0.103	0.053	0.229	0.031	-0.116
Replication similarity	5.72	1.05	6	3 to 7	0.015	-0.079	-0.052	-0.053	0.040
Difficulty of implementation	4.06	1.44	3	1 to 6	-0.072	0.000	-0.028	-0.108	-0.087
Replication df/N	12516	4575	71	5 to 768703	-0.085	-0.248	-0.696	-0.266	-0.136
Replication power	0.920	0.086	0.95	.56 to .99	0.374	-0.065	0.159	-0.032	0.330
Surprised by outcome	2.51	1.07	2	1 to 5	-0.468	0.332	-0.357	-0.371	-0.490

Appendices

[A1] Recalculation of *p*-values

Recalculation of p-values. The *p*-values were recalculated using the test statistic and the degrees of freedom, with the following R-function:

```
# Recalculating p-values
# Written by CHJ Hartgerink, RCM van Aert, MALM van Assen

pvalr <- function(x, N) {
  fis.r <- 0.5*log((1 + x) / (1 - x))
  se.fis.r <- sqrt(1/(N-3))
  pnorm(fis.r, mean = 0, sd = se.fis.r, lower.tail = FALSE)
}

# Computes two-tailed p-value
pvalComp <- function(
  x,
  df1,
  df2,
  N,
  esType){
  pvalComp <- ifelse(esType=="t",
                     pt(abs(x), df = df2, lower.tail = FALSE) * 2,
                     ifelse(
            esType=="F",
                     pf(x, df1 = df1, df2 = df2, lower.tail = FALSE),
                     ifelse(
                     esType=="r",
                     pvalr(abs(x), N) * 2,
                     ifelse(
                       esType=="Chi2",
                     pchisq(x, df = df1, lower.tail = FALSE),
                     ifelse(
                               esType == "z",
                               pnorm(abs(x), lower.tail = FALSE) * 2,
                          NA
                     )
                     )
                     )
                     ))

  return(pvalComp)
}
```

Remarks p-values and significance

- We used the 2-tailed recalculated p-values, with the exception of studies 7, 15, 47, 94, 120, 140 because the *p*-values were one-tailed (see column BU; *p*-values in DH and DT marked with yellow).
- For study 82 we used the reported *p*-value rather than the recalculated *p*-value, because there was a difference in test performed by the replication team (*t*-test for correlation) and the test used for recalculation (Fisher *z* test) that resulted in a different *p*-value (marked with green in column DT).
- For study 69 the *p*-values of both original and replication study were entered manually. In these studies, six highly significant binomial tests were carried out. We entered '.000001' in columns DH and DT for these studies (marked with green).
- The *p*-values of study 59 could neither be retrieved nor recalculated, although both are known to be significant (marked with purple in DH, DT, EA, and EB).
- Four *p*-values of original studies were interpreted as significant, although these *p*-values were larger than .05. When creating variables for the statistical significance of the effect (columns EA and EB), these effects were coded as significant (marked with red).

[A2] Analyses of significance and *p*-values

The code for the McNemar test of change in statistical significance:

```
#  McNemar test
tab <- table(dat$sign..O.[!is.na(dat$sign..O.) & !is.na(dat$sign..R.)],
             dat$sign..R.[!is.na(dat$sign..O.) & !is.na(dat$sign..R.)])
mcnemarchi <- (tab[1,2]-tab[2,1])^2/(tab[1,2]+tab[2,1])
mcnemarp <- pchisq(q = mcnemarchi, df = 1, lower.tail = FALSE)
```

The CIs of proportions of significance were computed exactly using the following TURBO Pascal routine:

```
PROGRAM confidence_for_p;
{$N+}
USES CRT;

CONST n = 5;
      ns = 5;
      a: array[1..5] of extended = (0.001,0.01,0.025,0.05,0.10);

var count: integer;
    h,p1,p2,po,pb: extended;
    co,cb: array[1..5] of extended;
    ob: integer;

{-----------------------------------------------------------------------}
function fac(i: integer): extended;
var c: integer;
    h: extended;

begin
    h:= 1;
    for c:= 1 to i
    do h:= h*c;
    fac:= h;
end;
{-----------------------------------------------------------------------}
function bin(n,i: integer): extended;
begin
    bin:= fac(n)/( fac(i)*fac(n-i) );
end;
{-----------------------------------------------------------------------}
function cdf(p: extended): extended;
var c: integer;
```

```
begin
   h:= 0;
   for c:= 0 to (ns-ob)
   do h:= h + bin(n,c) * exp( c*ln(p) ) * exp( (n-c)*ln(1-p) );
   cdf:= h;
end;
{------------------------------------------------------------------------}

begin
   p1:= 0;
   ob:= 1;
   if not (ns/n = 0)
   then begin
           for count:= 1 to 5
           do begin
                 p2:= ns/n;
                 if count > 1
                 then p1:= co[count-1];
                 repeat
                    po:= (p1+p2)/2;
                    if cdf(po) > 1-a[count]
                    then p1:= po
                    else p2:= po;
                 until abs(cdf(po)-1+a[count]) < 0.000001;
                 co[count]:= po;
              end;
        end;
   ob:= 0;
   p2:= 1;
   if not (ns/n = 1)
   then begin
            for count:= 1 to 5
            do begin
                  p1:= ns/n;
                  if count > 1
                  then p2:= cb[count-1];
                  repeat
                     po:= (p1+p2)/2;
                     if cdf(po) > a[count]
                     then p1:= po
                     else p2:= po;
                  until abs(cdf(po)-a[count]) < 0.000001;
                  cb[count]:= po;
               end;
        end;
end.
```

The code for the (Fisher, *p*-curve, *p*-uniform) test of no evidential value in the non-significant replication studies:

```
# Written by CHJ Hartgerink
# The Fisher method applied to test for deviation from uniformity
# In NONSIGNIFICANT P-values

FisherMethod <- function(# Compute Fisher's exact test for non-significant p-values.
	### This function computes paper level Fisher test statistics, testing whether the distribution of non-significant p-values is uniform. Significant values indicate deviation from uniformity.
	### Returns both the normal Fisher test, as well as the complement test.
	### Computations are done for p*=log(p), where p is all non-significant p-values for each identifier.
	x,
	### Vector of p-values.
	id,
	### Vector giving paper identifiers.
	alpha = .05
	### Indicate what alpha level is being maintained for the study results, which serves as a cut-off for selecting the non-significant p-values.
	){
	Res <- NULL
	for(i in 1:length(unique(id)))
	{
		selP <- x[id==unique(id)[i]]
		nSigP <- (na.omit(selP[selP>alpha])-alpha)/(1-alpha)
		SigP <- na.omit(selP[selP<=alpha])
		if(!length(nSigP)==0){
			# Compute the Fisher test statistic
			FMeth <- -2*sum(log(nSigP))
			# Compute p-values analytically
			pFMeth <- pchisq(q=FMeth, df=2*length(nSigP), lower.tail=F)
			} else {
				FMeth <- NA
				pFMeth <- NA
			}
		Res <- rbind(Res, data.frame(
			Fish = FMeth,
			PFish = pFMeth,
			CountNSig = length(nSigP),
			CountSig = length(SigP),
			PercentNonSig = length(nSigP)/length(selP)))
```

```
        }
        return(Res)
}
```

The code for the test comparing the means of the two dependent samples:

```
# Dependent t-test p-values
t.test(x = dat$pval_USE..O.[!is.na(dat$pval_USE..O.) & !is.na(dat$pval_USE..R.)],
        y = dat$pval_USE..R.[!is.na(dat$pval_USE..O.) & !is.na(dat$pval_USE..R.)],
        paired = TRUE)

# Wilcoxon signed-rank test p-values
wilcox.test(dat$pval_USE..O.[!is.na(dat$pval_USE..O.) & !is.na(dat$pval_USE..R.)],
         dat$pval_USE..R.[!is.na(dat$pval_USE..O.) & !is.na(dat$pval_USE..R.)],
        alternative="two.sided")

sd(dat$pval_USE..O.[!is.na(dat$pval_USE..O.) & !is.na(dat$pval_USE..R.)])
summary(dat$pval_USE..O.[!is.na(dat$pval_USE..O.) & !is.na(dat$pval_USE..R.)])
sd(dat$pval_USE..R.[!is.na(dat$pval_USE..O.) & !is.na(dat$pval_USE..R.)])
summary(dat$pval_USE..R.[!is.na(dat$pval_USE..O.) & !is.na(dat$pval_USE..R.)])
```

[A3] Calculation of effect sizes

Whenever possible, we calculated the "correlation coefficient per df" as effect size measure based on the reported test statistics. This was possible for the *z*, χ^2, *t*, and *F* statistic. The code for the calculation is:

```
esComp <- function(
  x,
  df1,
  df2,
  N,
  esType){
 esComp <- ifelse(esType=="t",
   sqrt((x^2*(1 / df2)) / (((x^2*1) / df2) + 1)),
   ifelse(
          esType=="F",
          sqrt((x*(df1 / df2)) / (((x*df1) / df2) + 1))*sqrt(1/df1),
          ifelse(
          esType=="r",
          x,
          ifelse(
                  esType=="Chi2",
                  sqrt(x/N),
                  ifelse(
                          esType == "z",
                          tanh(x * sqrt(1/(N-3))),
                          NA
                          )
                  )
          )
   ))
 return(esComp)
```

The *z* statistic is transformed into a correlation using sample size *N* with $z = r_f\sqrt{(N-3)}$, with r_f the Fisher-transformed correlation. The χ^2 is transformed into the or correlation coefficient with $\phi = \sqrt{\chi^2/N}$. The *t* and *F* statistic are transformed into a "correlation per *df*" using $r = \sqrt{\frac{F\frac{df_1}{df_2}}{F\frac{df_1}{df_2}+1}}\sqrt{\frac{1}{df_1}}$, where $F = t^2$. The expression in the first square-root equals the proportion of variance explained by the df_1 predictors of the variance not yet explained by these same predictors. To take into account that more predictors can explain more variance, we divided this number by df_1 to obtain the "explained variance by predictor". Taking the square root gives the correlation, or more precisely, it gives the correlation of each predictor

assuming that all df_1 predictors contribute equally to the explained variance of the dependent variable.

The correlation effect sizes can be found in columns DJ and DV of the master data file.

Remarks effect sizes

- The effect sizes of original studies 120 and 154 were marked with green (column DV), to indicate that both original and replication effect are coded as positive although their observed effects were negative.
- Seventeen studies are marked with orange because the sign of the replication effect was different from the sign of the original effect (column DV); in these cases, the original effect was (as always) coded as positive and the replication effect as negative.
- No “correlation per df” effect size could be computed for study-pairs 59, 69, and 77, hence 97 study-pairs have data on “correlation per df”.
- For study-pairs 59 and 69 effect sizes could be compared on another scale than the correlation (see columns BQ and BZ for 59, and BG and BZ for 69; marked with green in column EC). For study-pair 77 the effect sizes could not be compared (marked with red in column EC).
- The table below lists all effect sizes or test statistics (first column) and their frequency (second column), and for which analyses on comparisons of effect size they could be included (columns three to six). The last row presents the frequency of study-pairs for each of the analyses in the columns.

Effect size or test statistic	Frequency	% Comparison (which effect is larger?)	Meta-analytic estimate (subset MA)	% meta-analytic ($p < .05$)	% original effect size within replication 95% CI
t or *F*(1,*df*)	70	+	+	+	+
F(>1,*df*)	17	+	X	X	+
χ^2 odds ratio	2	+	X	+	+
χ^2 other	2	+	X	X	+
Binomial	1	+	X	X	X
r	4	+	+	+	+
beta and F	1	X	X	X	X
b	1	+	X	X	X
z	2	+	X	X	X
Total frequency	100	99	74	76	95

[A4] Calculation of expected coverage of original effect size by replication CI

One statistic to evaluate reproducibility is the probability that the original study's effect size is covered by the replication study's confidence interval. If $\alpha = .05$, and we assume that both studies are sampled from a population with the same true effect size, then this probability is a function of both studies' effect size. When both studies have equal sample size, this probability equals 83.4% (*39*). However, this probability can be any number between 0 (if the replication study has a much larger sample size) and 1 (if the original study has a much larger sample size).

The program below calculates the expected proportion of coverage across study pairs, by summing the study pairs' probabilities. For each study, the probability of overlap is calculated using the Fisher transformed effect size and its standard error. Since the standard error can only be calculated for test statistics *t*, *F*(1,df), and *r*, we can only use this statistic for study pairs who used these tests.

```
overlap <- numeric()
points <- 1000000
p <- 1:points/(points+1)        # uniform probability density based on equally distributed points

for (i in 1:length(final$N.r)) {
  zu <- qnorm(p,0,1/sqrt(final$N.r[i]-3)) + qnorm(.975)/sqrt(final$N.r[i]-3)
        # zu gives upper bound of Fisher transformed effect size for each possible point in the
        # probability density
  zl <- zu - 2*qnorm(.975)/sqrt(final$N.r[i]-3)
        # zl gives lower bound of Fisher transformed effect size for each possible point in the
        # probability density
  overlap[i] <- mean(pnorm(zu,0,1/sqrt(final$N.o[i]-3))) -
mean(pnorm(zl,0,1/sqrt(final$N.o[i]-3)))
        # overlap gives the probability of coverage as the average proportion that the original
effect
        # size is lower than the upper bound minus the average proportion that the original
effect
        # is larger than the lower bound
}
overlap
mean(overlap)
```

[A5] Calculation of expected coverage of original effect size by replication CI for other statistics

Effect size statistics based on $F(df_1 > 1, df_2)$ and $\chi^2(df)$ can be converted to correlations (see A3), but their standard errors cannot be computed. Hence, coverage, or the probability that the original study's effect size is covered by the replication study's confidence interval, needs to be computed in another way. For *F* statistics we first computed the 95% confidence interval of the non-centrality parameter based on the observed *F*-statistic of the replication study. Then, we estimated the non-centrality parameter λ of the original study using the fact that the expected value of the *F*-statistic equals

$$F = \frac{df_1 + \lambda}{df_1} \times \frac{df_2}{df_2 - 2}.$$

Rewriting this expected value yields $\hat{\lambda} = \frac{(df_2 - 2) \times F \times df_1}{df_2}$. Coverage then means that the CI contains $\hat{\lambda}$. See the code below.

Similarly, for $\chi^2(df)$ statistics we checked if the CI of the non-centrality parameter λ of the replication study contains the estimated non-centrality parameter $\hat{\lambda}$ of the original study. Using the fact that the expected value of the non-central chi-square distribution equals $df + \lambda$, we obtain $\hat{\lambda}$ = $\chi^2(df)$ -1, which $\chi^2(df)$ equal to the test statistic of the original study. The CI contains $\hat{\lambda}$, if the cumulative probability of the chi-square value of the replication study given $\hat{\lambda}$ is between .025 and .975. See the code below.

```
tol <- 1e-7
xm <- 0

df1.or <- df2.or <- F.or <- df1.rep <- df2.rep <- F.rep <- 1:17
ncp.L <- ncp.U <- ncp.o <- in.ci <- 1:17

### study 12
df1.or[1] <- 2
df2.or[1] <- 92
F.or[1] <- 3.13
df1.rep[1] <- 2
df2.rep[1] <- 232
F.rep[1] <- 1.63

### study 13
df1.or[2] <- 2
df2.or[2] <- 68
F.or[2] <- 41.59
df1.rep[2] <- 2
```

```
df2.rep[2] <- 68
F.rep[2] <- 41.603

### study 17
df1.or[3] <- 2
df2.or[3] <- 76
F.or[3] <- 8.67
df1.rep[3] <- 1.58
df2.rep[3] <- 72.4
F.rep[3] <- 19.48

### study 22
df1.or[4] <- 3
df2.or[4] <- 93
F.or[4] <- 5.23
df1.rep[4] <- 2.33
df2.rep[4] <- 90
F.rep[4] <- 0.38

### study 43
df1.or[5] <- 2
df2.or[5] <- 64
F.or[5] <- 10.17
df1.rep[5] <- 2
df2.rep[5] <- 72
F.rep[5] <- 1.97

### study 46
df1.or[6] <- 21
df2.or[6] <- 230025
F.or[6] <- 118.15
df1.rep[6] <- 21
df2.rep[6] <- 455304
F.rep[6] <- 261.93

### study 50
df1.or[7] <- 2
df2.or[7] <- 92
F.or[7] <- 4.36
df1.rep[7] <- 2
df2.rep[7] <- 103
F.rep[7] <- 2.601
```

```
### study 55
df1.or[8] <- 2
df2.or[8] <- 54
F.or[8] <- 3.19
df1.rep[8] <- 2
df2.rep[8] <- 68
F.rep[8] <- 0.3

### study 64
df1.or[9] <- 2
df2.or[9] <- 76
F.or[9] <- 21.57
df1.rep[9] <- 2
df2.rep[9] <- 65
F.rep[9] <- 0.865

### study 80
df1.or[10] <- 2
df2.or[10] <- 43
F.or[10] <- 3.36
df1.rep[10] <- 2
df2.rep[10] <- 67
F.rep[10] <- 1.7

### study 86
df1.or[11] <- 2
df2.or[11] <- 82
F.or[11] <- 4.05
df1.rep[11] <- 2
df2.rep[11] <- 137
F.rep[11] <- 1.99

### study 117
df1.or[12] <- 18
df2.or[12] <- 660
F.or[12] <- 16.31
df1.rep[12] <- 18
df2.rep[12] <- 660
F.rep[12] <- 12.98

### study 132
df1.or[13] <- 3
df2.or[13] <- 69
```

```
F.or[13] <- 5.15
df1.rep[13] <- 1.48
df2.rep[13] <- 41.458
F.rep[13] <- 1.401

### study 139
df1.or[14] <- 3
df2.or[14] <- 9
F.or[14] <- 8.5
df1.rep[14] <- 3
df2.rep[14] <- 12
F.rep[14] <- 13.06

### study 140
df1.or[15] <- 2
df2.or[15] <- 81
F.or[15] <- 4.97
df1.rep[15] <- 2
df2.rep[15] <- 122
F.rep[15] <- 0.24

### study 142
df1.or[16] <- 2
df2.or[16] <- 162
F.or[16] <- 192.89
df1.rep[16] <- 2
df2.rep[16] <- 174
F.rep[16] <- 252.83

### study 143
df1.or[17] <- 4
df2.or[17] <- 108
F.or[17] <- 3.67
df1.rep[17] <- 4
df2.rep[17] <- 150
F.rep[17] <- 0.58

### loop
for (i in 1:length(F.or)) {
  df1.o <- df1.or[i]
  df2.o <- df2.or[i]
  F.o <- F.or[i]
  df1.r <- df1.rep[i]
```

```
  df2.r <- df2.rep[i]
  F.r <- F.rep[i]

  ### ncp lower bound
  if (pf(F.r,df1.r,df2.r,0) < .975)
  {ncp.L[i] <- 0} else
  {
          x0 <- 0
          x1 <- df1.r*F.r
          print(x1)
          ym <- 1
          while(abs(ym-0.975) > tol) {
          xm <- (x0+x1)/2
          ym <- pf(F.r,df1.r,df2.r,xm)
          if (ym > 0.975) x0 <- xm
          if (ym < 0.975) x1 <- xm
          print(xm)
          print(ym)
          }
          ncp.L[i] <- xm
  }

  ### ncp upper bound
  x0 <- df1.r*F.r
  x1 <- 20*df1.r*F.r
  print(x0)
  print(x1)
  ym <- 1
  while(abs(ym-0.025) > tol) {
          xm <- (x0+x1)/2
          ym <- pf(F.r,df1.r,df2.r,xm)
          if (ym > 0.025) x0 <- xm
          if (ym < 0.025) x1 <- xm
          print(xm)
  }
  ncp.U[i] <- xm

  ### check if original is in ci of replication
  ncp.o[i] <- F.o*df1.o*(df2.o-2)/df2.o-df1.o
  in.ci[i] <- ( (ncp.L[i] < ncp.o[i]) & (ncp.U[i] > ncp.o[i]) )
}

cbind(ncp.L,ncp.o,ncp.U,in.ci)
```

```
sum(in.ci)
mean(in.ci)

### ch2
## if probability calculated with pchisq is between .025
## and .975 then the ncp of original is in ci of replication

## Study 73
chi2.o <- 3.85
chi2.r <- 4.8
pchisq(chi2.r,1,chi2.o-1)

## Study 84
chi2.o <- 13.18
chi2.r <- 7.1
pchisq(chi2.r,1,chi2.o-1)

## Study 104
chi2.o <- 3.83
chi2.r <- 0.387
pchisq(chi2.r,1,chi2.o-1)

## Study 165
chi2.o <- 4.51
chi2.r <- 1.57
pchisq(chi2.r,1,chi2.o-1)
```

[A6] Analyses of Effect Sizes

The code for the first two tests comparing means of dependent samples:

```
# Dependent t-test effects (r values)
t.test(x = dat$r..O.[!is.na(dat$r..O.) & !is.na(dat$r..R.)],
       y = dat$r..R.[!is.na(dat$r..O.) & !is.na(dat$r..R.)],
      paired = TRUE)

# Wilcox test effects (r values)
wilcox.test(dat$r..O.[!is.na(dat$r..O.) & !is.na(dat$r..R.)],
       dat$r..R.[!is.na(dat$r..O.) & !is.na(dat$r..R.)],
        alternative="two.sided")

summary(dat$r..O.[!is.na(dat$r..O.) & !is.na(dat$r..R.)])
sd(dat$r..O.[!is.na(dat$r..O.) & !is.na(dat$r..R.)])
summary(dat$r..R.[!is.na(dat$r..O.) & !is.na(dat$r..R.)])
sd(dat$r..R.[!is.na(dat$r..O.) & !is.na(dat$r..R.)])

mean(dat$r..O.[!is.na(dat$r..O.) & !is.na(dat$r..R.)])-mean(dat$r..R.[!is.na(dat$r..O.) &
!is.na(dat$r..R.)])
```

The third test comparing effect sizes ('which is stronger?') was carried out using the variable comparing effect sizes (column EC). The frequency of studies where the original effect size exceeded the replication effect size (*f*) and the total number of comparisons (*n*) were entered in the binomial test:

```
binom.test(f, n, 0.5, "two.sided", 0.95)
```

The fourth and last test compared the observed proportion of study-pairs in which the effect of the original study was in the confidence interval of the effect of the replication study with the expected proportion using a goodness-of-fit χ^2-test. Supplement [A4] provides the code for calculating the expected proportion. The code for calculating the observed proportion can be found in supplement [A6]. The observed frequency *f*, expected proportion *p*, and number of comparisons *n* was entered in the binomial test:

```
binom.test(f, n, p, "two.sided", 0.95)
```

The number of comparisons *n* equals the number of studies in which the effect was tested using *r*, *t*, or *F*(1,df).

[A7] Meta-analyses on effect sizes of each study-pair

The meta-analyses were conducted on Fisher-transformed correlations for all study-pairs in subset MA, i.e. for all study-pairs where both the correlation coefficient and its standard error could be computed. Standard errors could only be computed if test statistics were *r*, *t*, or $F(1,df_2)$, which was for 74 study-pairs. Standard errors of Fisher-transformed correlations were computed using $1/\sqrt{df_2 - 1}$, which assumes tests of one correlation or an independent sample *t*-test (but not a dependent sample *t*-test).

The results of all individual meta-analyses are reported after the code.

```
################################
### Meta-analyses per pair ###
################################

### How often is the null hypotheses rejected in the meta-analysis
in.ci <- es.meta <- se.meta <- ci.lb.meta <- ci.ub.meta <- pval.meta <- numeric()

for(i in 1:length(final$fis.o)) {
  tmp <- rma(yi = c(final$fis.o[i], final$fis.r[i]), sei = c(final$sei.o[i], final$sei.r[i]), method = "FE")
  es.meta[i] <- tmp$b[1]
  se.meta[i] <- tmp$se
  ci.lb.meta[i] <- tmp$ci.lb
  ci.ub.meta[i] <- tmp$ci.ub
  pval.meta[i] <- tmp$pval

  if(tmp$pval < 0.05) { in.ci[i] <- 1
  } else { in.ci[i] <- 0 }
}

sum(in.ci)/length(in.ci) # Proportion of times the null hypothesis of no effect is rejected

### Create data frame
tab <- data.frame(ID = final$ID, fis.o = final$fis.o, sei.o = final$sei.o, pval.o = final$pval.o, fis.r
= final$fis.r, sei.r = final$sei.r,
                  pval.r = final$pval.r, diff = final$yi, es.meta = es.meta, se.meta = se.meta,
ci.lb.meta = ci.lb.meta, ci.ub.meta = ci.ub.meta, pval.meta = pval.meta)

### Check how often effect size original study is within CI of meta-analysis
in.ci.meta <- numeric()

for(i in 1:length(final$fis.o)) {
```

```
  if(final$fis.o[i] > ci.lb.meta[i] & final$fis.o[i] < ci.ub.meta[i]) {
    in.ci.meta[i] <- TRUE
  } else { in.ci.meta[i] <- FALSE }

}

sum(in.ci.meta)/length(in.ci.meta) # Proportion of times the original study is within the CI of
meta-analysis

#################################################################
### How often is original study within CI of replication ###
#################################################################

### Create confidence interval for replications
ci.lb <- final$fis.r-qnorm(.975)*final$sei.r
ci.ub <- final$fis.r+qnorm(.975)*final$sei.r

in.ci <- numeric()

for(i in 1:length(final$fis.r)) {

  if (final$fis.o[i] > ci.lb[i] & final$fis.o[i] < ci.ub[i]) {
          in.ci[i] <- TRUE
  } else { in.ci[i] <- FALSE }

}

sum(in.ci)/length(in.ci) # Proportion of times the original study is within the CI of the replication
```