

1
2
3 Many Labs 5: Testing pre-data collection peer review as an intervention to increase replicability
4
5

6 Charles R. Ebersole, Department of Psychology, University of Virginia

7
8 Maya B. Mathur, Quantitative Sciences Unit, Stanford University

9
10 Erica Baranski, University of Houston

11
12 Diane-Jo Bart-Plange, Department of Psychology, University of Virginia

13
14 Nicholas R. Buttrick, Department of Psychology, University of Virginia

15
16 Christopher R. Chartier, Department of Psychology, Ashland University

17
18 Katherine S. Corker, Grand Valley State University

19
20 Martin Corley, Psychology, PPLS, University of Edinburgh

21
22 Joshua K. Hartshorne, Department of Psychology, Boston College

23
24 Hans IJzerman, LIP/PC2S, Université Grenoble Alpes

25
26 Ljiljana B. Lazarevic, Institute of Psychology and Laboratory for Research of Individual Differences, University of
27 Belgrade

28
29 Hugh Rabagliati, Psychology, PPLS, University of Edinburgh

30
31 Ivan Ropovik, Charles University, Faculty of Education, Institute for Research and Development of Education &
32 University of Presov, Faculty of Education

33
34 Balazs Aczel, Institute of Psychology, Eötvös Loránd University, Hungary

35
36 Lena F. Aeschbach, Department of Psychology, University of Basel

37
38 Luca Andrighetto, Department of Educational Science, University of Genova, Italy

39
40 Jack D. Arnal, McDaniel College

41
42 Holly Arrow, Department of Psychology, University of Oregon

43
44 Peter Babincak, Institute of Psychology, Faculty of Arts, University of Presov

45
46 Bence E. Bakos, Institute of Psychology, Eötvös Loránd University, Hungary

47
48 Gabriel Baník, Institute of Psychology, Faculty of Arts, University of Presov

49
50 Ernest Baskin, Department of Food Marketing, Haub School of Business, Saint Joseph's University

51
52 Radomir Belopavlović, Department of Psychology, University of Novi Sad, Serbia

53
54 Michael H. Bernstein, Center for Alcohol and Addiction Studies, Brown University; Department of Psychology,
55 University of Rhode Island

1
2
3
4 Michał Białek, Department of Economic Psychology, Kozminski University, Poland

5 Nicholas G. Bloxson, Department of Psychology, Ashland University

6
7 Bojana Bodroža, Department of Psychology, Faculty of Philosophy, University of Novi Sad, Serbia

8
9 Diane B. V. Bonfiglio, Department of Psychology, Ashland University

10
11 Leanne Boucher, Department of Psychology and Neuroscience, Nova Southeastern University

12
13 Florian Brühlmann, Department of Psychology, University of Basel

14
15 Claudia Brumbaugh, Department of Psychology, The Graduate Center and Queens College, City University of New
16 York

17
18 Erica Casini, University of Milano - Bicocca, Italy

19
20 Yiling Chen, John A. Paulson School of Engineering and Applied Sciences, Harvard University

21
22 Carlo Chiorri, Department of Educational Science, University of Genova, Italy

23
24 William J. Chopik, Department of Psychology, Michigan State University

25
26 Oliver Christ, Fernuniversität in Hagen, Germany

27
28 Antonia M. Ciunci, Department of Psychology, University of Rhode Island

29
30 Heather M. Claypool, Department of Psychology, Miami University

31
32 Sean Coary, Department of Food Marketing, Haub School of Business, Saint Joseph's University

33
34 Marija V. Čolić, Faculty of Sport and Physical Education, University of Belgrade, Serbia

35
36 W. Matthew Collins, Department of Psychology and Neuroscience, Nova Southeastern University

37
38 Paul G. Curran, Department of Psychology, Grand Valley State University

39
40 Chris R. Day, Centre for Trust, Peace and Social Relations, Coventry University, UK

41
42 Benjamin Dering, Psychology, University of Stirling, UK

43
44 Anna Dreber, Department of Economics, Stockholm School of Economics, Sweden, and Department of Economics,
45 University of Innsbruck, Austria

46
47 John E. Edlund, Rochester Institute of Technology

48
49 Filipe Falcão, Department of Psychology, University of Porto, Portugal

50
51 Anna Fedor, MTA-ELTE Theoretical Biology and Evolutionary Ecology Research Group, Budapest, Hungary

52
53 Lily Feinberg, Department of Psychology, Boston College

54
55 Ian R. Ferguson, Department of Psychology, Virginia Commonwealth University

1
2
3
4 Máire Ford, Department of Psychology, Loyola Marymount University

5 Michael C. Frank, Department of Psychology, Stanford University

6
7 Emily Fryberger, Department of Psychology, Pacific Lutheran University

8
9 Alexander Garinther, Department of Psychology, University of Oregon

10
11 Katarzyna Gawryluk, Department of Economic Psychology, Kozminski University, Poland

12
13 Kayla Gerken, Rose-Hulman Institute of Technology

14
15 Mauro Giacomantonio, Department of Social & Developmental Psychology, Sapienza University of Rome

16
17 Steffen R. Giessner, Rotterdam School of Management, Erasmus University, The Netherlands

18
19 Jon E. Grahe, Department of Psychology, Pacific Lutheran University

20
21 Rosanna E. Guadagno, Center for International Security and Cooperation, Stanford University

22
23 Ewa Hałasa, Maria Curie-Skłodowska University, Poland

24
25 Peter J.B. Hancock, Psychology, University of Stirling, UK

26
27 Rias A. Hilliard, Rose-Hulman Institute of Technology

28
29 Joachim Hüffmeier, Department of Psychology, TU Dortmund University, Germany

30
31 Sean Hughes, Department of Experimental-Clinical and Health Psychology, Ghent University

32
33 Katarzyna Idzikowska, Department of Economic Psychology, Kozminski University, Poland

34
35 Michael Inzlicht, Department of Psychology, University of Toronto

36
37 Alan Jern, Rose-Hulman Institute of Technology

38
39 William Jiménez-Leal, Department of Psychology, Universidad de los Andes

40
41 Magnus Johannesson, Department of Economics, Stockholm School of Economics, Sweden

42
43 Jennifer A. Joy-Gaba, Department of Psychology, Virginia Commonwealth University

44
45 Mathias Kauff, Medical School Hamburg, Germany

46
47 Danielle J. Kellier, Perelman School of Medicine, University of Pennsylvania

48
49 Grecia Kessinger, Department of Psychology, Brigham Young University- Idaho

50
51 Mallory C. Kidwell, Department of Psychology, University of Utah

52
53 Amanda M. Kimbrough, College of Art, Technology, and Emerging Communication, University of Texas at Dallas

54
55 Josiah P. J. King, Psychology, PPLS, University of Edinburgh

1
2
3
4 Vanessa S. Kolb, Department of Psychology, University of Rhode Island

5 Sabina Kołodziej, Department of Economic Psychology, Kozminski University, Poland

6
7 Marton Kovacs, Institute of Psychology, Eötvös Loránd University, Hungary

8
9 Karolina Krasuska, Maria Curie-Skłodowska University, Poland

10
11 Sue Kraus, Psychology, Fort Lewis College, Durango, Colorado

12
13 Lacy E. Krueger, Texas A&M University-Commerce

14
15 Katarzyna Kuchno, Maria Curie-Skłodowska University, Poland

16
17 Caio Ambrosio Lage, Department of Psychology, Pontifical Catholic University of Rio de Janeiro, Brazil

18
19 Eleanor V. Langford, Department of Psychology, University of Virginia

20
21 Carmel A. Levitan, Department of Cognitive Science, Occidental College

22
23 Tiago Jessé Souza de Lima, Department of Social and Work Psychology, University of Brasília, Brazil

24
25 Hause Lin, Department of Psychology, University of Toronto

26
27 Samuel Lins, Department of Psychology, University of Porto, Portugal

28
29 Jia E. Loy, LEL, PPLS, University of Edinburgh

30
31 Dylan Manfredi, Marketing Department, The Wharton School of Business, University of Pennsylvania

32
33 Łukasz Markiewicz, Department of Economic Psychology, Kozminski University, Poland

34
35 Madhavi Menon, Department of Psychology and Neuroscience, Nova Southeastern University

36
37 Brett Mercier, Department of Psychological Science, University of California Irvine

38
39 Mitchell Metzger, Department of Psychology, Ashland University

40
41 Venus Meyet, Department of Psychology, Brigham Young University- Idaho

42
43 Ailsa E. Millen, Psychology, University of Stirling, UK

44
45 Jeremy K. Miller, Department of Psychology, Willamette University

46
47 Andres Montealegre, Universidad de los Andes

48
49 Don A. Moore, University of California at Berkeley

50
51 Rafał Muda, Maria Curie-Skłodowska University, Poland

52
53 Gideon Nave, Marketing Department, The Wharton School of Business, University of Pennsylvania

54
55 Austin Lee Nichols, Department of Business, University of Navarra, Spain

56

57

58

59

60

1
2
3
4 Sarah A. Novak, Department of Psychology, Hofstra University

5
6 Christian Nunnally, Rose-Hulman Institute of Technology

7
8 Ana Orlić, Faculty of Sport and Physical Education, University of Belgrade, Serbia

9
10 Anna Palinkas, Eötvös Loránd University

11 Angelo Panno, Department of Education, Experimental Psychology Laboratory, Roma Tre University

12
13 Kimberly P. Parks, Department of Psychology, University of Virginia

14
15 Ivana Pedović, Department of Psychology, University of Niš, Serbia

16
17 Emilian Pȩkala, Maria Curie-Skłodowska University, Poland

18
19 Matthew R. Penner, Department of Psychological Sciences, Western Kentucky University

20
21 Sebastiaan Pessers, University of Leuven, Belgium

22
23 Boban Petrović, Institute of Criminological and Sociological Research, Serbia

24
25 Thomas Pfeiffer, New Zealand Institute for Advanced Study, Massey University, New Zealand

26
27 Damian Pieńkosz, Maria Curie-Skłodowska University, Poland

28
29 Emanuele Preti, University of Milano - Bicocca, Italy

30
31 Danka Purić, Department of Psychology and Laboratory for Research of Individual Differences, University of
32 Belgrade, Serbia

33
34 Tiago Ramos, Department of Psychology, University of Porto, Portugal

35
36 Jonathan Ravid, Department of Psychology, Boston College

37
38 Timothy S. Razza, Department of Psychology and Neuroscience, Nova Southeastern University

39
40 Katrin Rentzsch, Department of Psychology, University of Goettingen, Germany, and Leibniz Science Campus
41 Primate Cognition

42
43 Juliette Richetin, University of Milano-Bicocca, Italy

44
45 Sean C. Rife, Murray State University

46
47 Anna Dalla Rosa, Department of Philosophy, Sociology, Education and Applied Psychology, University of Padova

48
49 Kaylis Hase Rudy, Department of Psychology, Brigham Young University- Idaho

50
51 Janos Salamon, Doctoral School of Psychology, Eötvös Loránd University, Institute of Psychology, Eötvös Loránd
52 University, Hungary

53
54 Blair Saunders, Psychology, School of Social Sciences, University of Dundee

55
56 Przemysław Sawicki, Department of Economic Psychology, Kozminski University, Poland

1
2
3
4 Kathleen Schmidt, Department of Psychology, Southern Illinois University Carbondale

5
6 Kurt Schuepfer, Department of Psychology, Miami University

7
8 Thomas Schultze, Department of Psychology, University of Goettingen, Germany, and Leibniz Science Campus
9 Primate Cognition

10
11 Stefan Schulz-Hardt, Department of Psychology, University of Goettingen, Germany, and Leibniz Science Campus
12 Primate Cognition

13
14 Astrid Schütz, Department of Psychology, University of Bamberg, Germany

15
16 Ani Shabazian, Loyola Marymount University, USA

17
18 Rachel L. Shubella, Rose-Hulman Institute of Technology

19
20 Adam Siegel, Cultivate Labs

21
22 Rúben Silva, Department of Psychology, University of Porto, Portugal

23
24 Barbara Sioma, Maria Curie-Skłodowska University, Poland

25
26 Lauren Skorb, Department of Psychology, Boston College

27
28 Luana Elayne Cunha de Souza, University of Fortaleza, Brazil

29
30 Sara Steegen, University of Leuven, Belgium

31
32 LAR Stein, Psychology Department, University of Rhode Island; Center for Alcohol and Addiction Studies and
33 Department of Behavioral & Social Sciences, Brown University.

34
35 R. Weylin Sternglanz, Department of Psychology and Neuroscience, Nova Southeastern University

36
37 Darko Stojilović, Department of Psychology, University of Belgrade, Serbia

38
39 Daniel Storage, Department of Psychology, University of Denver

40
41 Gavin Brent Sullivan, Centre for Trust, Peace and Social Relations, Coventry University, UK

42
43 Barnabas Szaszi, Institute of Psychology, Eötvös Loránd University, Hungary

44
45 Peter Szecsi, Institute of Psychology, Eötvös Loránd University, Hungary

46
47 Orsolya Szoke, Institute of Psychology, Eötvös Loránd University, Hungary

48
49 Attila Szuts, Institute of Psychology, Eötvös Loránd University, Hungary

50
51 Manuela Thomae, Department of Psychology, University of Winchester, United Kingdom

52
53 Natasha D. Tidwell, Department of Psychology, Fort Lewis College, Durango, CO

54
55 Carly Tocco, Department of Psychology, The Graduate Center, City University of New York, New York, New York
56 and Department of Psychology, Queens College, City University of New York, Flushing, NY

57
58
59
60

1
2
3
4 Ann-Kathrin Torke, Department of Psychology, TU Dortmund University, Germany

5 Francis Tuerlinckx, University of Leuven, Belgium

6
7 Wolf Vanpaemel, University of Leuven, Belgium

8
9 Leigh Ann Vaughn, Department of Psychology, Ithaca College

10
11 Michelangelo Vianello, Department of Philosophy, Sociology, Education and Applied Psychology, University of
12 Padova

13
14 Domenico Viganola, Department of Economics, Stockholm School of Economics, Sweden

15
16 Maria Vlachou, University of Leuven, Belgium

17
18 Ryan J. Walker, Department of Psychology, Miami University

19
20 Sophia C. Weissgerber, Universität Kassel, Germany

21
22 Aaron L. Wichman, Psychological Sciences Department, Western Kentucky University

23
24 Bradford J. Wiggins, Department of Psychology, Brigham Young University - Idaho

25
26 Daniel Wolf, Department of Psychology, University of Bamberg, Germany

27
28 Michael J. Wood, Department of Psychology, University of Winchester, United Kingdom

29
30 David Zealley, Department of Psychology, Brigham Young University - Idaho

31
32 Iris Žeželj, Department of Psychology and Laboratory for Research of Individual Differences, University of
33 Belgrade, Serbia

34
35 Mark Zrubka, Eötvös Loránd University, Hungary

36
37 Brian A. Nosek, Center for Open Science and Department of Psychology, University of Virginia

Abstract

Replications in psychological science sometimes fail to reproduce prior findings. If replications use methods that are unfaithful to the original study or ineffective in eliciting the phenomenon of interest, then a failure to replicate may be a failure of the protocol rather than a challenge to the original finding. Formal pre-data collection peer review by experts may address shortcomings and increase replicability rates. We selected 10 replications from the Reproducibility Project: Psychology (RP:P; Open Science Collaboration, 2015) in which the original authors had expressed concerns about the replication designs before data collection and only one of which was “statistically significant” ($p < .05$). Commenters suggested that lack of adherence to expert review and low-powered tests were the reasons that most of these RP:P studies failed to replicate (Gilbert et al., 2016). We revised the replication protocols and received formal peer review prior to conducting new replications. We administered the RP:P and Revised protocols in multiple laboratories (Median number of laboratories per original study = 6.5; Range 3 to 9; Median total sample = 1279.5; Range 276 to 3512) for high-powered tests of each original finding with both protocols. Overall, Revised protocols produced similar effect sizes as RP:P protocols following the preregistered analysis plan ($\Delta r = .002$ or $.014$, depending on analytic approach). The median effect size for Revised protocols ($r = .05$) was similar to RP:P protocols ($r = .04$) and the original RP:P replications ($r = .11$), and smaller than the original studies ($r = .37$). The cumulative evidence of original study and three replication attempts suggests that effect sizes for all 10 (median $r = .07$; range $.00$ to $.15$) are 78% smaller on average than original findings (median $r = .37$; range $.19$ to $.50$), with very precisely estimated effects.

Total words = 289

Keywords = replication, reproducibility, metascience, peer review, Registered Reports

Many Labs 5: Testing pre-data collection peer review as an intervention to increase replicability

The replicability of evidence for scientific claims is important for scientific progress. The accumulation of knowledge depends on reliable past findings to generate new ideas and extensions that can advance understanding. Not all findings will replicate -- researchers will inevitably later discover that some findings were false leads. However, if problems with replicability are pervasive and unrecognized, scientists will struggle to build on previous work to generate cumulative knowledge and will have difficulty constructing effective theories.

Large-sample, multi-study replications have failed to replicate a substantial portion of the published findings that they tested. For example, based on each of their primary replication criterion, success rates include: Klein et al. (2014) 10 of 13 findings (77%) successfully replicated; Open Science Collaboration (2015) 36 of 97 (37%)¹; Camerer et al. (2016) 11 of 18 (61%); Ebersole et al. (2016) 3 of 10 (30%); Cova et al. (2018) 29 of 37 (78%); Camerer et al. (2018) 13 of 21 (62%); and Klein et al. (2018) 14 of 28 (50%). Moreover, replications, even when finding supporting evidence for the original claim (e.g., $p < .05$) tend to show a smaller observed effect size compared to the original study. For example, Camerer et al. (2018) successfully replicated 13 of 21 social science studies originally published in the journals *Science* and *Nature*, but the average effect size of the successful replications was only 75% of the original and the average effect size of the unsuccessful replications was near zero. These studies are not a random sample of social-behavioral research, but the cumulative evidence suggests that there is room for improvement, particularly for a research culture that has not historically prioritized publishing replications (Makel et al., 2012).

¹ RP:P included 100 replications, however 3 of the original studies showed null results.

1
2
3 A finding might not replicate for several reasons. The initial finding might have been a
4 false positive, reflecting either a “normal” Type I error or one made more likely through
5 selective reporting of positive results and ignoring null results (Greenwald, 1975; Rosenthal,
6 1979; Sterling, 1959), or by employing flexibility in analytic decisions and reporting (Gelman &
7 Loken, 2014; John et al., 2012; Simmons, Nelson, & Simonsohn, 2011). Alternatively, the theory
8 being tested might be insufficiently developed, such that it cannot anticipate possible moderators
9 inadvertently introduced in the replication study (Simons, Shoda, & Lindsay, 2017). Finally, the
10 replication study might have been a false negative, reflecting either a lack of statistical power or
11 an ineffective or unfaithful methodology that disrupted detecting the true effect. Many prior
12 replication efforts attempted to minimize false negatives by using large samples, obtaining
13 original study materials, and requesting feedback from original authors on study protocols before
14 they were administered. Nevertheless, these design efforts may not have been sufficient to
15 reduce or eliminate false negatives for true effects. For example, in the Reproducibility Project:
16 Psychology (RP:P; Open Science Collaboration, 2015), replication teams sought materials and
17 feedback from original authors to maximize the quality of the 100 replication protocols. In 11
18 cases, studies were identified as “not endorsed” meaning that original authors had identified
19 potential shortcomings *a priori* that were not addressed in the ultimate design.² These
20 shortcomings may have had implications for replication success. Of the 11 studies, only 1
21 successfully replicated the original finding, albeit much more weakly than the original study.
22 These unresolved issues were cited in a critique of RP:P as a likely explanation for replication
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52

53 ² There has been some confusion over the procedure for labeling endorsement of RP:P studies (e.g., Gilbert, King,
54 Pettigrew, & Wilson, 2016). Assessments of original author endorsement were made by replication teams prior to
55 conducting the replication. They assessed what they believed the authors’ endorsement to be, based on whether or
56 not the replication design had addressed any concerns raised by the original authors.
57
58
59
60

1
2
3 failure (Gilbert, King, Pettigrew, & Wilson, 2016; but see responses by Anderson et al., 2016;
4
5 Nosek & Gilbert, 2016).

6 7 **Unfaithful or Ineffective Methods as a Moderator of Replicability**

8
9
10 Replication is attempting to reproduce a previously observed finding with no *a priori*
11
12 expectation for a different outcome (see Nosek & Errington, 2020; Nosek & Errington, 2017;
13
14 Zwaan et al., 2018). Nevertheless, a replication may still produce a different outcome for a
15
16 variety of reasons (Gilbert, King, Pettigrew, & Wilson, 2016; Luttrell, Petty, & Xu, 2017; Noah,
17
18 Schul, & Mayo, 2018; Open Science Collaboration, 2015; Petty & Cacioppo, 2016; Stroebe &
19
20 Strack, 2014; Schwarz & Strack, 2014; Strack, 2016). Replicators could fail to implement key
21
22 features of the methodology that are essential for observing the effect. They could also
23
24 administer the study to a population for which the finding is not expected to apply. Alternatively,
25
26 replicators could implement features of the original methodology that are not appropriate for the
27
28 new context of data collection. For example, in a study for which object familiarity is a key
29
30 feature, objects familiar to an original sample in Europe might not be similarly familiar to a new
31
32 sample in Asia. A more appropriate test of the original question might require selecting new
33
34 objects that have comparable familiarity ratings across populations (e.g., Chen, Chartier, &
35
36 Szabelska, 2018, replications of Stanfield & Zwaan, 2001). These simultaneous challenges of (a)
37
38 adhering to the original study, and (b) adapting to the new context, have the important
39
40 implication that claims over whether or not a particular study is a replication is theory-laden
41
42 (Nosek & Errington, 2017; 2020). Since exact replication is impossible, claiming “no *a priori*
43
44 expectation for a different outcome” is an assertion that all of the differences between the
45
46 original study and the replication are theoretically irrelevant for observing the identified effect.
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 Like all theoretical claims, asserting that a new study is a replication of a prior study
4 cannot be proven definitively. In most prior large-scale replication projects, replication teams
5 made final decisions about study protocols after soliciting feedback from original authors or
6 other experts. Such experts may be particularly well-positioned to assess weaknesses in study
7 protocols and their applicability to new circumstances for data collection. Despite genuine efforts
8 to solicit and incorporate such feedback, insufficient attention to expert feedback may be part of
9 the explanation for existing failures to replicate (Gilbert et al., 2016).
10
11
12
13
14
15
16
17
18

19 The studies in RP:P that were “not endorsed” by original authors offer a unique
20 opportunity to test this hypothesis. The RP:P protocols were deemed by the replication teams to
21 be replications of the original studies. However, original authors expressed concerns prior to data
22 collection. Thus, if any failed replications can be explained due to poor replication design, these
23 are among the top candidates. Thus, we revised 10 of the 11 “non-endorsed” protocols from
24 RP:P and subjected them to peer review before data collection, a model known as Registered
25 Reports (Chambers, 2013; Nosek & Lakens, 2014; <http://cos.io/rr/>). Once the protocols were
26 accepted following formal peer review, they were preregistered on OSF (see Table 1). Then, we
27 conducted replications using both the RP:P protocols and the Revised protocols, with multiple
28 laboratories contributing data for one or both protocols. This “many labs” design allowed us to
29 achieve unusually high statistical power, decreasing the probability that any failure to replicate
30 could be due to insufficient power.
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46

47 This design is particularly well-suited for testing the strong hypothesis that many, if not
48 most, failures to replicate are due to design errors that could have been caught by a domain
49 expert (Gilbert et al., 2016). If this hypothesis is correct, then the new, peer-reviewed protocol
50 should improve replicability and increase effect sizes to be closer to the original studies. This
51
52
53
54
55
56
57
58
59
60

1
2
3 would not necessarily mean that *all* failures to replicate are due to poor design -- our sample of
4 studies was chosen because they are among the most likely published replications to have faulty
5 designs -- but it would suggest that published replicability rates are overly pessimistic. Note that
6 the replications using the original RP:P protocols serve as a control: If both protocols lead to
7 successful replications, then the failures in RP:P were more likely due to low power or some
8 unexpected difference in the replication teams themselves. In contrast, if most of the replications
9 fail even after expert input, it casts doubt on the “design error” hypothesis, at least for these
10 studies. Rather, such an outcome would increase the likelihood that the original findings were
11 false positives because even formal expert input had no effect on improving replicability.
12
13
14
15
16
17
18
19
20
21
22
23

24 Finally, in parallel with the replication attempts, we organized a group of independent
25 researchers to participate in surveys and prediction markets to bet on whether the RP:P and
26 Revised protocols would successfully replicate the original findings. Prior evidence suggests that
27 researchers can effectively anticipate replication success or failure with surveys and prediction
28 markets (Camerer et al., 2016; Camerer et al., 2018; Dreber et al., 2015; Forsell et al., 2018). As
29 such, this provided an opportunity to test whether researchers anticipated improvements in
30 replicability between the RP:P and Revised protocols and whether those predictions were related
31 to actual replication success. If so, it might suggest that design errors and potential for improving
32 replicability can be predicted *a priori* through markets or surveys.
33
34
35
36
37
38
39
40
41
42
43

44 **Disclosures**

45
46
47 Confirmatory analyses were preregistered on OSF (<https://osf.io/nkmc4/>). Links to the
48 preregistrations for the individual replications can be found in Table 1. All materials, data, and
49 code are available on the OSF (<https://osf.io/7a6rd/>). The RP:P Protocols were created from the
50 original RP:P materials that can be found here: <https://osf.io/ezcuj/>. We report how we
51
52
53
54
55
56
57
58
59
60

1
2
3 determined our sample size, all data exclusions, all manipulations, and all measures in the study.
4
5 Data were collected in accordance with the Declaration of Helsinki. The authors acknowledge a
6
7 conflict-of-interest that Brian Nosek is Executive Director of the non-profit Center for Open
8
9 Science that has a mission to increase openness, integrity, and reproducibility of research. This
10
11 project was supported by a grant from the Association for Psychological Science and from
12
13 Arnold Ventures. In addition, the following authors thank other sources of funding: the French
14
15 National Research Agency (ANR-15-IDEX-02; IJzerman), the Netherlands Organization for
16
17 Scientific Research (NWO) (016.145.049; IJzerman), the National Institute on Alcohol Abuse
18
19 and Alcoholism (F31AA024358; MH Bernstein), the Social Sciences and Humanities Research
20
21 Council of Canada (149084; Inzlicht), the Economic and Social Research Council (UK,
22
23 ES/L01064X/1, Rabagliati), John Templeton Foundation (Ebersole and Nosek), Templeton
24
25 World Charity Foundation (Nosek), and Templeton Religion Trust (Nosek). The authors thank
26
27 the many original authors and experts who provided extensive feedback throughout the many
28
29 stages of the project. CRE and BAN conceived the project and drafted the report. MBM and
30
31 CRE designed the analysis plan and analyzed the aggregate data. CRE, CRC, JKH, HIJ, IR,
32
33 MBM, LBL, HR, MC, EB, DB, KSC, and NRB served as team leaders for the sets of
34
35 replications. DV, CRE, YC, TP, AD, MJ, AS, and BAN designed and analyzed the surveys and
36
37 prediction markets to elicit peer beliefs. All authors except BAN collected the data. All authors
38
39 revised and approved the manuscript with two exceptions; sadly, Sebastiaan Pessers and Boban
40
41 Petrović passed away before the manuscript was finalized.
42
43
44
45
46
47
48

49 **Method**

50 **Replications**

51
52
53
54
55
56
57
58
59
60

1
2
3 Our selection criteria for studies to replicate consisted of those labeled “not-endorsed”
4 from RP:P (Open Science Collaboration, 2015). For each of the 11 candidate studies, we sought
5 team leads to conduct the new replications and enough research teams to satisfy our sampling
6 plan (see below). We recruited researchers through professional listservs, personal contacts, and
7 collaboration websites (StudySwap, <https://osf.io/view/StudySwap/>). We were able to satisfy our
8 recruitment goals for 10 of the 11 replications (all except Murray et al., 2008). For each of the 10
9 studies, we conducted two replications with multiple samples each: one using the RP:P protocol,
10 and the other using the Revised protocol that was approved following formal peer review.
11 Because RP:P focused on a single statistical result from each original study, both protocols
12 focused on replicating that same result.
13
14
15
16
17
18
19
20
21
22
23
24
25

26 **Preparation of Protocols and Peer-Review**

27
28 Teams reconstructed each RP:P protocol using the methods and materials that were
29 shared by the original RP:P replication teams (<https://osf.io/ezcuuj/>). This protocol was the basis
30 for the RP:P protocol condition. Differences between the RP:P Protocol and the replication as
31 described in RP:P reflected practicalities such as lab space, population, climate, and time of year
32 (see other articles, this issue, for details of those replications). Next, teams sought out any
33 correspondence and/or responses written by the original authors concerning the RP:P
34 replications.³ Teams revised the RP:P protocols to account for concerns expressed in those
35 sources. This revision was the basis for the Revised protocol condition. Then, both the RP:P
36 protocols and the Revised protocols were submitted for peer-review through *Advances in*
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

³ Correspondence from RP:P (OSC, 2015) was accessed from that project’s OSF page (osf.io/ezcuuj/).

1
2
3 authors were unavailable or unwilling to provide a review, the Editor sought input from other
4
5 experts. Based on editorial feedback, teams updated their Revised protocols and resubmitted
6
7 them for additional review until the protocols were given in-principle acceptance.
8
9

10 The peer review process produced a range of requested revisions across replication
11
12 studies. Some revisions concerned using a participant sampling frame more similar to that of the
13
14 original study (e.g., some RP:P protocols differed from original studies regarding sampling from
15
16 the lab vs. MTurk, different countries, different age ranges). Some revisions increased
17
18 methodological alignment of the Revised protocol with that of the original study. Other revisions
19
20 altered the protocol from the original to make it more appropriate for testing the original research
21
22 question in the replication contexts. Importantly, we were agnostic to which types of changes
23
24 would be most likely to yield successful replications. We sought to enact all revisions that
25
26 experts deemed important to make successful replication as likely as possible and that were
27
28 feasible given available resources. If there were disagreements about the feasibility of a request,
29
30 the Editor made a final decision.
31
32
33
34

35 Upon acceptance, teams preregistered their protocols on OSF and initiated data
36
37 collection. Table 1 provides links to the preregistered protocols and brief summaries of the main
38
39 differences between the RP:P and Revised protocols (e.g., the primary changes to the protocol
40
41 suggested by reviewers/previous correspondence). The reports for each of the 10 studies were
42
43 submitted for results-blind review so that the Editor and reviewers could examine how
44
45 confirmatory analyses would be conducted and presented. To ensure that the authors and
46
47 reviewers could discuss the current study's methods and analysis plan without being biased by
48
49 the results, the present summary report was drafted and peer-reviewed prior to the two project
50
51 organizers knowing the results of the majority of the replications (BAN knew none of the results;
52
53
54
55
56
57
58
59
60

1
2
3 CRE was directly involved with data collection for two of the sets of replications and was aware
4 of only those results). CRE and BAN had primary responsibility for drafting the paper, and all
5 other authors contributed to revisions. Other authors knew outcomes of 0 or 1 of the sets of
6 replications during the writing process depending on which individual studies they helped
7 conduct. The full reports of each individual replication are reported separately in this issue
8 [CITATIONS TO BE ADDED]. All data, materials, code, and other supplementary information
9 are available at <https://osf.io/7a6rd/>.

19 **Sampling Plan**

20
21 We collected data for 20 protocols in total -- 2 versions (RP:P and Revised) for each of
22 10 original studies.⁴ For each protocol, we sought a minimum of 3 data collection sites unless the
23 study sampled from MTurk (i.e., the RP:P protocol of Risen & Gilovich, 2008). At each site, we
24 sought a sample that achieved 95% power to detect the effect size reported in the original study
25 ($\alpha = .05$). If we expected that the target sample size would be difficult to collect at every site, we
26 recruited additional collection sites for that protocol so that the test based on the total sample size
27 would be highly powered. Overall, samples in this project (RP:P protocols: mean $N = 805.20$,
28 median $N = 562.5$, $SD = 787.82$; Revised protocols: mean $N = 590.30$, median $N = 629.50$, $SD =$
29 391.72) were larger than those of the original studies (mean $N = 70.8$, median $N = 76$, $SD =$
30 34.25) and RP:P replications (mean $N = 103$, median $N = 85.5$, $SD = 61.94$). We calculated
31 power to detect the original effect size with $\alpha = .05$ for each of the protocols. Overall, our studies
32 were very well powered to detect the original effect sizes (see Table 2). When possible, we
33 randomly assigned participants to one protocol or the other within each data collection site. This

53
54 ⁴ The replication of van Dijk et al. (2008) included an additional, Web-based protocol. This was motivated by a
55 desire to test certain predictions made by the original authors. However, because it matches neither the RP:P
56 protocol nor what was recommended during review, it is not included in the analysis here. For more detail, see
57 Skorb et al. (this issue).

1
2
3 was possible for half of the studies; for the other half randomization was impossible due to the
4
5 revisions to the RP:P protocol (e.g., MTurk vs. in-lab collection).
6
7

8 **Eliciting peer beliefs**

9
10 Predictions about replication success guided the selection and revision of studies to
11
12 replicate in this project. To assess whether other researchers shared these predictions, we
13
14 measured peer beliefs about the replications. Following previous efforts (Dreber et al., 2015;
15
16 Camerer et al., 2016; 2018; Forsell et al., 2018), we invited psychology researchers to predict the
17
18 replication outcomes for the 10 RP:P protocols and 10 Revised protocols in prediction markets
19
20 and surveys. Before being allowed to trade in the markets, participants had to rate the probability
21
22 of the binary measure of successful replication (a statistically significant effect at $p < 0.05$ in the
23
24 same direction as the original study) for each of the 20 protocols in a survey. In the prediction
25
26 market, participants traded contracts worth money if the study replicated and worth nothing if the
27
28 study did not replicate. With some caveats (Manski, 2006), the prices of such contracts can be
29
30 interpreted as the probabilities that the market assign the studies replicating. For each study,
31
32 participants could enter the quantity of the contract they wanted to buy (if they believed that the
33
34 true probability that the study will replicate is higher than the one specified by the current price)
35
36 or to sell (if they believed that the true probability that the study will replicate is lower than the
37
38 one identified by the current price). Participants were endowed with points corresponding to
39
40 money that we provided, and they thus had a monetary incentive to report their true beliefs. For
41
42 each study, participants were provided with links to the RP:P protocols, the Revised protocols,
43
44 and to a document summarizing the differences between the two. They were informed that all the
45
46 replications had a power of at least 80%. The prediction markets were open for two weeks
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 starting from June 21st, 2017, and a total of 31 participants made at least one trade. See the
4
5 Supplemental Material for more details about the prediction markets and survey.
6
7

8 **Power Analyses**

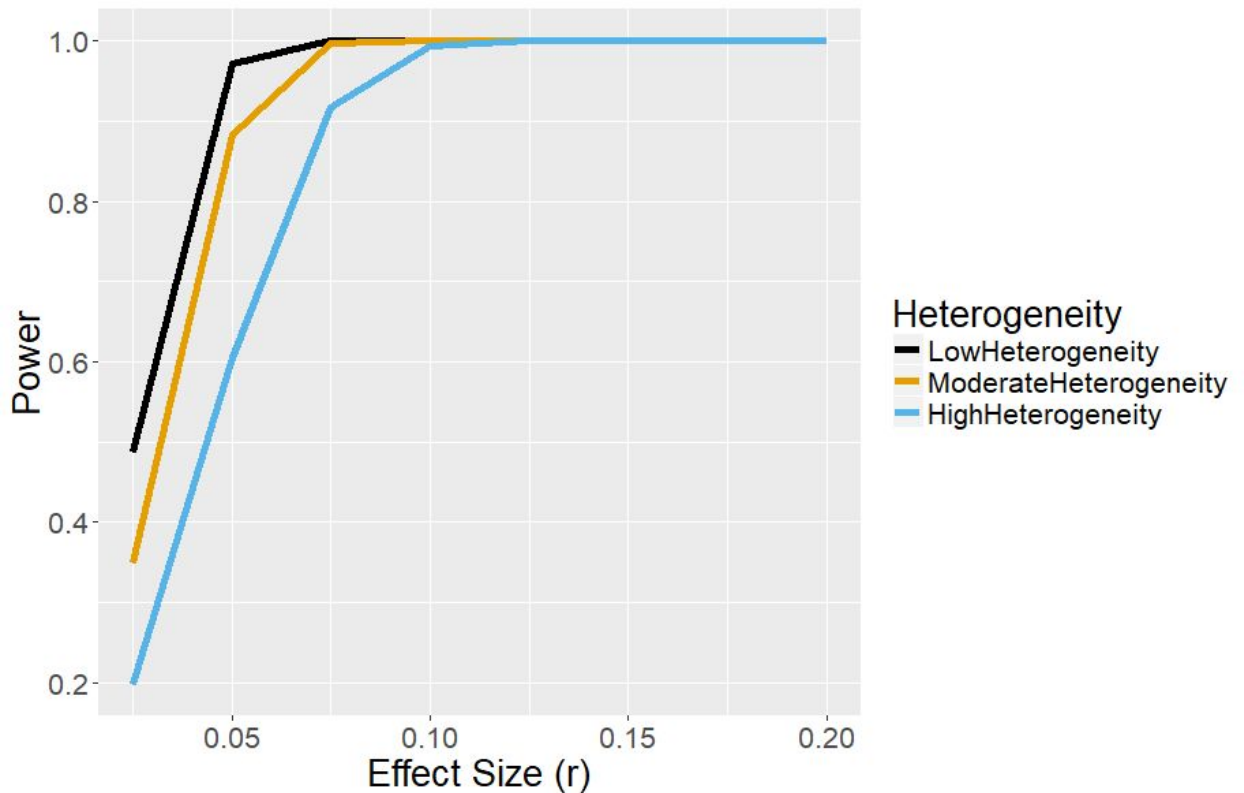
9
10 The primary test for this study involved comparing the replicability of studies using
11 protocols from RP:P compared to those using protocols revised through expert peer review. We
12 calculated our power to detect such an effect, measured as the effect of protocol within each set
13 of studies ($k = 10$). The results are displayed in Figure 1.⁵ In cases of both low ($I^2 = 25\%$) and
14 moderate ($I^2 = 50\%$) heterogeneity, our minimum planned samples should provide adequate
15 power ($> 80\%$) to detect an average effect of protocol as small as $r = .05$. For greater
16 heterogeneity ($I^2 = 75\%$), our minimum planned samples should provide adequate power to
17 detect an effect of protocol as small as $r = .075$. Power under all heterogeneity assumptions
18 approaches 100% for effects of $r = .10$ or larger. As a comparison, the difference between effect
19 sizes reported in the original studies and those reported in RP:P were, on average, $\Delta r = .27$.
20
21
22
23
24
25
26
27
28
29
30
31
32

33 We also simulated our estimated power for a second analysis strategy, that being meta-
34 analyzing the effect sizes from each protocol within each individual site and testing protocol
35 version as a meta-analytic moderator.⁶ These power estimates were slightly lower.
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54

55 ⁵ See <https://osf.io/j5vnh/> for power and figure script.

56 ⁶ See <https://osf.io/dhr3p/> for power simulation script.
57
58
59
60

Figure 1 - Power to detect effect of protocol



At relatively high heterogeneity ($I^2 = 73-75\%$), our minimum planned sample would achieve adequate power at an average effect size difference between protocols of $\Delta r = .125$ (90% power). However, it is worth noting that both sets of power analyses rely on making assumptions about the amount of different sources of heterogeneity. The observed heterogeneity will be informative for understanding the sensitivity of these tests.

Finally, we estimated power for detecting relationships between peer beliefs and replication outcomes. The twenty prediction markets would provide 41% power to detect a correlation of 0.4, 62% power to detect a correlation of 0.5, 82% power to detect a correlation of 0.6, and 95% power to detect a correlation of 0.7. The previous prediction markets have found an

1
2
3 average correlation of 0.58 between peer beliefs and replication outcomes (78% power with
4
5 twenty markets).
6

7 8 **Results**

9 10 **Confirmatory Analyses - Comparing Results from RP:P and Revised Protocols**

11
12 We replicated 10 studies with two large-sample protocols, one based on the RP:P
13
14 replication study (Open Science Collaboration, 2015), and the other that was Revised based on
15
16 formal peer review by experts. In the original papers, all ten key findings were statistically
17
18 significant ($p < .05$), the median effect size magnitude was $r = .37$, and the median sample size
19
20 was $N = 76$. In RP:P, 1 of 10 findings was statistically significant ($p < .05$), the median effect
21
22 size was $r = .11$, and the median sample size was $N = 85.5$.
23
24
25

26
27 In the present study, 0 of 10 replications using the RP:P protocol yielded a “statistically
28
29 significant” meta-analytic effect size ($p < .05$), the median effect size was $r = .04$, and the
30
31 median sample size was $N = 562.5$. Also in the present study, 2 of 10 replications⁷ using the
32
33 Revised protocol yielded statistically significant meta-analytic effect sizes ($p < .05$), the median
34
35 effect size was $r = .07$, and the median sample size was $N = 629.5$. Gauging replication success
36
37 based on whether the replications are statistically significant is subject to substantial caveats. For
38
39 example, depending on the power of the original study and replications, the expected proportion
40
41 of significant replications can be quite low even when the original is consistent with the
42
43 replications (Andrews & Kasy, 2019; Patil, Peng, & Leek, 2016). Therefore, as a benchmark to
44
45 help interpret these metrics regarding statistical significance, we estimated the expected
46
47 probability that each pooled replication estimate would be “statistically significant” and positive
48
49
50
51
52
53

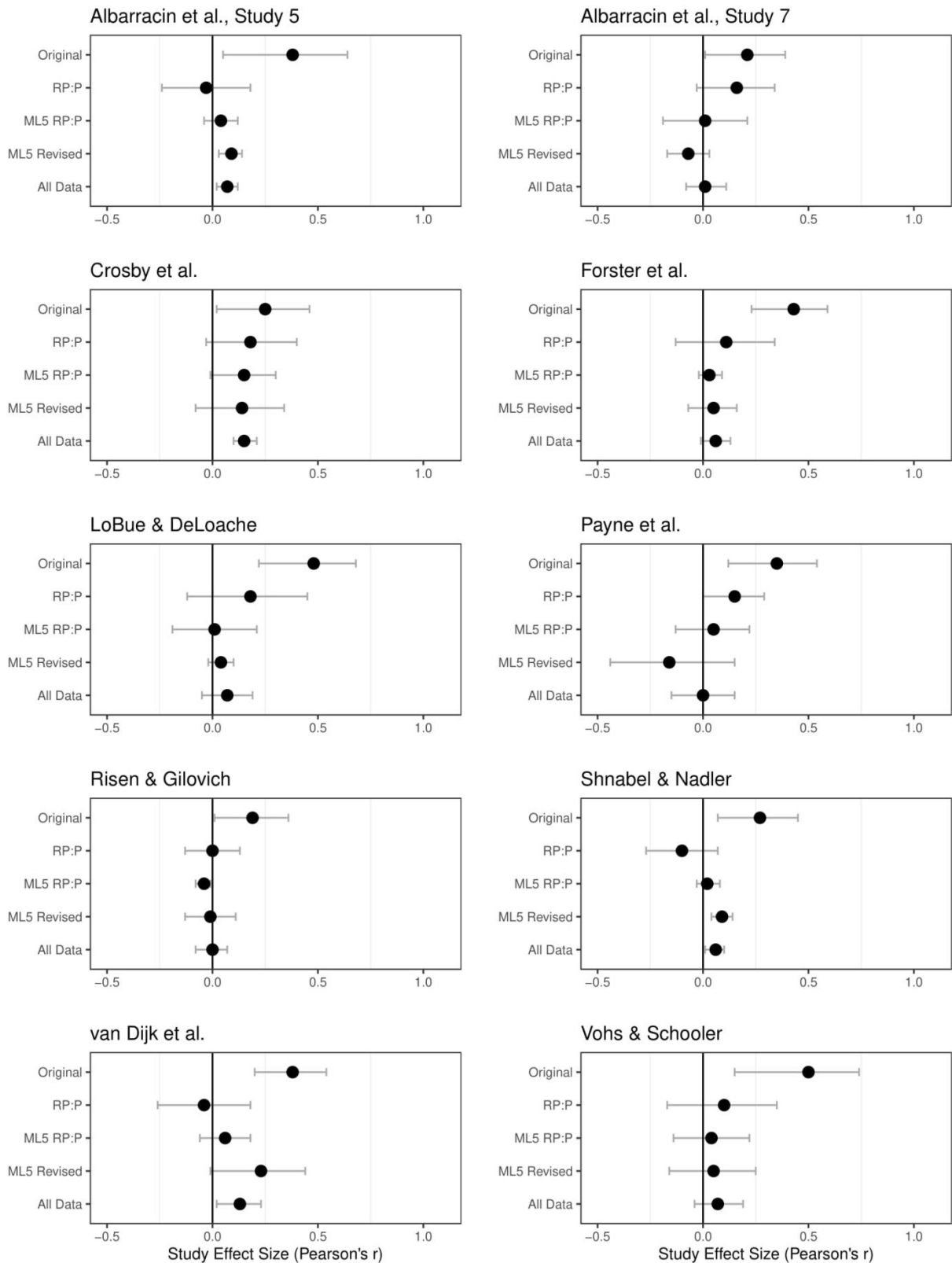
54
55 ⁷ The results in this paper focus on meta-analytic outcomes across the 10 pairs of studies. The individual reports of
56
57 the 10 pairs of studies tended to use mixed-effects models to gauge the statistical significance of each replication.
58
59 The statistical significance of each protocol in each study may vary as a result.
60

1
2
3 in sign, if in fact the replications were consistent with the original study (Mathur &
4
5 VanderWeele, 2020). A full summary of aggregated effect sizes and confidence intervals for
6
7 each data collection appears in Table 3.
8
9

10 The purpose of this investigation was to test whether a protocol resulting from formal
11
12 peer review would produce stronger evidence for replicability than a protocol that had not
13
14 received formal peer review. We tested this in two ways. First, we calculated an effect size for
15
16 each protocol within each data collection site. Each site implementing both the RP:P protocol
17
18 and the Revised protocol contributed two effect sizes, and each site implementing only one of the
19
20 two protocols contributed one effect size. We conducted a multilevel random-effects meta-
21
22 analysis of the $k = 101$ effect sizes⁸, with a random intercept of data collection Site (varying from
23
24 3 to 9 depending on Study) nested within Study (10 studies). This model converged so we did
25
26 not alter the model further. Then, we added the protocol version (RP:P vs. Revised), the
27
28 hypothesized moderator, as a fixed effect. We found that it had a near zero effect, $b = .002$, $SE =$
29
30 $.02$, $z = .091$, $p = .928$, 95% CI [-.04, .04]. That is, effect sizes from Revised protocols were, on
31
32 average, $b = .002$ units on the Pearson's r scale larger than effect sizes from RP:P protocols.
33
34 Overall, effect sizes had little variance accounted for by the moderator as indexed by $\text{Tau} = .05$
35
36 (95% CI [0, .09]) on the Fisher's z scale. There was, however, significant heterogeneity between
37
38 the effect sizes overall, as indicated by the Q statistic, $Q = 147.07$, $p = .001$, $I^2 = 26.57\%$.
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54

55 ⁸ Throughout, we meta-analyzed effect sizes on the Fisher's z scale, but report results transformed back to the
56
57 Pearson's r scale for interpretability except where otherwise noted.
58
59
60

Figure 2 - Effect sizes across study versions



Note: "All Data" represents a random effects meta-analytic estimate including the original study, the RP:P replication (OSC, 2015), and all Many Labs 5 data.

1
2
3 For the second test, we conducted a random-effects meta-analysis on the estimates of the
4 effect of protocol within each replication. We calculated the strength of the effect of protocol on
5 the Pearson's r scale for each of the 10 studies. A meta-analysis of these $k = 10$ estimates
6 suggested that these effect sizes were not reliably different from zero, $b = .014$, $SE = .01$, $t =$
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

.968, $p = .335$, 95% CI [-.02, .05]. Across studies, the Revised protocol point estimates were thus
on average $b = .014$ units larger than the RP:P point estimate on the Pearson's r scale. Overall,
the effect of protocol within each study had a fairly small amount of heterogeneity as indicated
by $\text{Tau} = .034$ (95% CI [0, .06]) on the Fisher's z scale. However, the Q statistic suggested
significant heterogeneity, $Q = 21.81$, $p = .010$, $I^2 = 60.89\%$. Examining heterogeneity by study
(e.g., collapsing across protocols), only one of the individual studies showed at least a small
amount of heterogeneity as estimated by Tau being greater than 0.10: Payne et al. ($\text{Tau} = .16$)

Exploratory Analyses - Other Evaluations of Replicability

Both of our primary tests of the effect of formal peer review on increasing effect sizes of
replications failed to reject the null hypothesis and showed very weak effect sizes with narrow
confidence intervals. Nevertheless, 2 of the Revised protocols showed effects below the $p < .05$
threshold (p -values .009 and .005) while none of the RP:P protocols did so. While this pattern
might appear to support the hypothesis that expert peer review could improve replicability, vote
counting the number of "significant" replications is not a formal test (Mathur & VanderWeele,
2020). This pattern could have occurred by chance, and indeed the formal statistical tests do not
suggest that the difference is systematic. Perhaps formal peer review does not improve
replicability of findings more than trivially, but perhaps it did for these two studies? Of the two
statistically significant effects with the Revised protocol, the observed effect sizes were 76% and
67% smaller than the findings in the original paper. Comparing the RP:P and Revised protocols

1
2
3 for each of these findings, as was done in the individual reports for this project, indicates that for
4
5 only one of the two tests was the revised protocol effect size significantly larger (Albarracin et
6
7 al., Study 5, $p = .601$; Shnabel & Nadler, $p = .012$). Therefore, even looking at the most
8
9 promising examples of the effect of formal peer review on increasing replicability fails to
10
11 provide reliable support. It is possible that the expert feedback did reliably improve the Shnabel
12
13 and Nadler effect size, but given the number of tests, it is also plausible that this difference
14
15 occurred by chance.
16
17
18

19 We also examined the cumulative evidence for each of the 10 findings. Figure 2 shows
20
21 the combined evidence of the original study, RP:P replication, and both protocols from the
22
23 current investigation. The combined evidence provides the highest powered test to detect small
24
25 effects, and the most precise estimates. Four of the 10 showed a statistically significant effect,
26
27 though the observed effect sizes (median $r = .10$) were much smaller than the original papers'
28
29 effect sizes (median $r = .38$), and all highest bounds of the 95% confidence intervals were below
30
31 $r = .25$, most far below.
32
33
34

35 **Exploratory Analyses - Additional Measures of Replicability**

36
37 As exploratory analyses, we considered several other measures of replicability that
38
39 directly assess: (1) statistical consistency between the replications and the original studies; and
40
41 (2) the strength of evidence provided by the replications for the scientific effect under
42
43 investigation (Mathur & VanderWeele, 2020). These analyses also account for potential
44
45 heterogeneity in replications and for observed sample sizes in both the replications and the
46
47 original studies. Accounting for these sources of variability avoids potentially misleading
48
49 conclusions regarding replication success that can arise from metrics that do not account for
50
51 these sources of variability, such as agreement in statistical significance.
52
53
54
55
56
57
58
59
60

1
2
3 First, an original study can be considered statistically “consistent” with a set of
4
5 replications if the original study and the replications came from the same distribution of
6
7 potentially heterogeneous effects – that is, if the original study was not an anomaly (Mathur &
8
9 VanderWeele, 2020). We assessed statistical consistency using the metric P_{orig} . Analogous to a p -
10
11 value for the null hypothesis of consistency, this metric characterizes the probability that the
12
13 original study would have obtained a point estimate at least as extreme as was observed, if in fact
14
15 the original study were consistent with the replications. P_{orig} thus assesses whether the replications
16
17 were similar to those of the original study with small values of P_{orig} indicating less similarity and
18
19 larger values indicating more similarity.
20
21
22
23

24 Second, we assessed the strength of evidence provided by the replications for each
25
26 scientific hypothesis investigated in the original studies (Mathur & VanderWeele, 2020).
27
28 Specifically, we estimated the percentage of population effects, among the potentially
29
30 heterogeneous distribution from which the replications are a sample, that agree in direction with
31
32 the original study. This metric is generous toward the scientific hypothesis by treating all effects
33
34 in the same direction as the original study, even those of negligible size, as evidence in favor of
35
36 the hypothesis. More stringently, we also estimated the percentage of population effects that not
37
38 only agreed in direction with the original, but were also meaningfully strong by two different
39
40 criteria (i.e., $r > .10$ or $r > .20$). These metrics together assess whether the replications provided
41
42 standalone evidence for the scientific hypothesis, regardless of the estimate of the original study
43
44 itself.
45
46
47
48

49 For each study, we conducted these analyses for three subsets: 1) all replications,
50
51 regardless of which protocol they used; 2) replications using the RP:P protocol; and 3)
52
53 replications using the Revised protocol. Note that the three percentage metrics should be
54
55
56
57
58
59
60

1
2
3 interpreted cautiously for subsets of fewer than 10 replications that also have heterogeneity
4 estimates greater than 0, and we conducted sensitivity analyses excluding four such studies from
5 aggregated statistics (Mathur & VanderWeele, 2020; see Supplement for methodological
6 details). For replication subsets that had a heterogeneity estimate of exactly 0 or which had only
7 1 replication, we simply report the percentage as either 100% or 0% depending on whether the
8 single point estimate was above or below the chosen threshold.
9

10
11
12
13
14
15
16
17 Table 4 aggregates these results, showing the mean value of P_{orig} and the mean
18 percentages of effects stronger than $r = 0, .1, \text{ and } .2$ respectively. Despite our close
19 standardization of protocols across sites, 40% of replication sets within each of the 3 subsets had
20 heterogeneity estimates greater than 0, highlighting the importance of estimating heterogeneity
21 when assessing replications. Regarding statistical consistency between the originals and the
22 replications, the median values of P_{orig} were .04 and .02 for the Revised replications and the RP:P
23 replications, respectively. That is, there were on average 4% and 2% probabilities that the
24 original studies' estimates would have been at least as extreme as observed if, for each study, the
25 original and replication studies had come from the same distribution. Of Revised and RP:P
26 replications, 50% and 80% respectively provided fairly strong evidence for inconsistency with
27 the original study ($P_{\text{orig}} < .05$), and 20% and 30% respectively provided strong evidence for
28 inconsistency ($P_{\text{orig}} < .01$). Thus, both the Revised and the RP:P replications often suggested
29 statistical inconsistency with the original study, even after accounting for effect heterogeneity
30 and other sources of statistical variability.⁹ However, heuristically, evidence for inconsistency
31 might have been somewhat less pronounced in the Revised than in the RP:P replications.
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52

53
54 ⁹ We performed sensitivity analyses that excluded the replication subsets that had fewer than 10
55 replications as well as heterogeneity estimates greater than 0. In these analyses, the median values of P_{orig}
56 were .08 and .01 for the Revised replications and the RP:P replications, respectively. Of Revised and
57
58
59
60

1
2
3 Regarding evidence strength for the scientific hypotheses, for the Revised replications, on
4 average only 50% of population effects agreed in direction with the original study (as expected if
5 the average effect size were exactly zero), 20% were above a modest effect size of $r = .10$, and
6 10% were above $r = .20$. For the RP:P replications, on average, 60% of effects agreed in
7 direction with the original study, 10% were above $r = .10$, and 0% were above $r = .20$. These
8 results suggest that even after accounting for heterogeneity, the large majority of population
9 effects were negligibly small regardless of protocol version.¹⁰ Thus, in both the Revised and the
10 RP:P replications, the population effects did not reliably support the scientific hypotheses even
11 when generously considering all effects that agreed in direction with the original study as
12 providing support; furthermore, only a small minority of population effects in each case were
13 meaningfully strong in size.

24 Peer Beliefs

25 We tested to what extent prediction markets and surveys could successfully predict the
26 replication outcomes. Thirty-five people participated in the survey and, of these, 31 made at least
27 one trade on the prediction markets. All survey results are based on the participants that made at
28 least one trade.¹¹

29 The survey and prediction markets produce a collective peer estimate of the replication
30 success probability for each replication. The mean predicted probability of a statistically
31 significant replication was .286 (range .124-.591) for the 10 RP:P protocols and .296 (range

32 RP:P replications, 20% and 86% respectively had $P_{\text{orig}} < .05$, and 20% and 29% respectively had $P_{\text{orig}} < .01$.

33 ¹⁰ In sensitivity analyses as described in the previous footnote, in the Revised replications, we estimated
34 that 100%, 40%, and 20% of effects were stronger than $r = 0$, $r = .1$, and $r = .2$ respectively. In the RP:P
35 replications, we estimated that 86%, 14%, and 0% of effects were stronger than these thresholds.

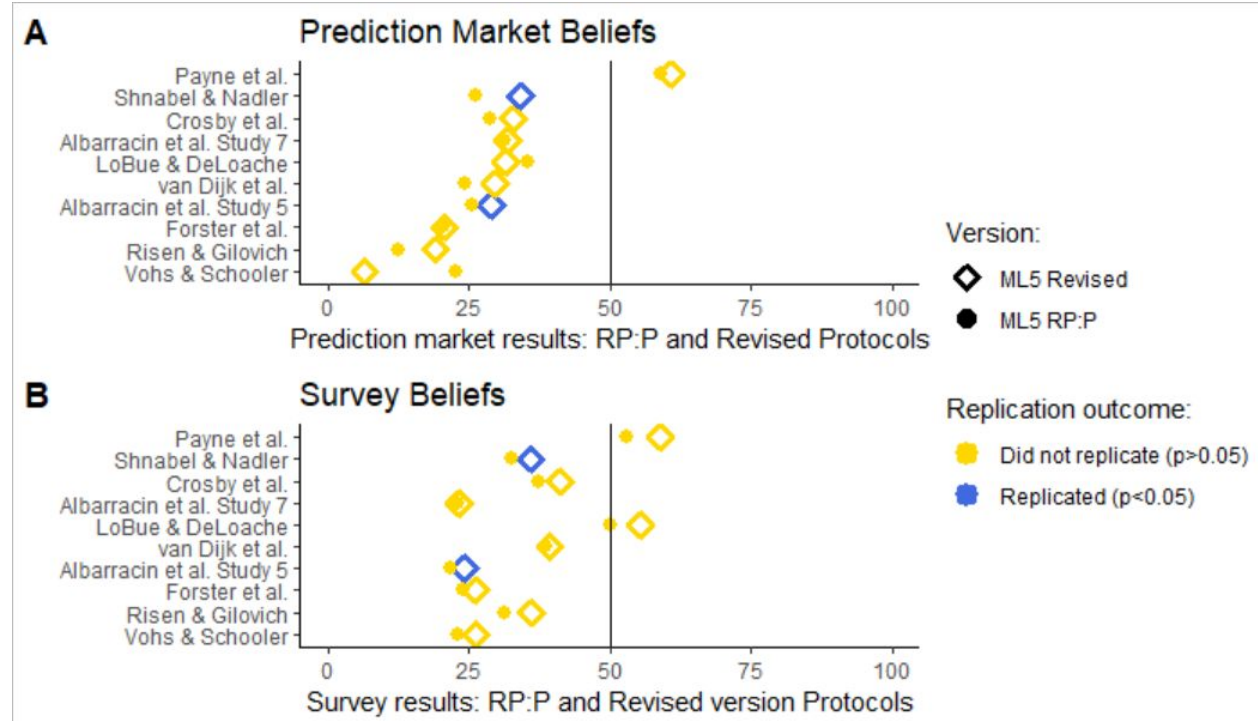
36 ¹¹ Many Labs 5 contributors were not allowed to make predictions on their studies, and their answers in
37 the survey about those studies were not used.

1
2
3 .065-.608) for the 10 Revised protocols (Wilcoxon signed-rank test, $p = .232$, $n = 10$), implying
4 that participants expected about 3 of 10 studies from each protocol to replicate. The mean survey
5 belief about replication success was .335 (range .217-.528) for the 10 RP:P protocols and .367
6 (range .233-.589) for the 10 Revised protocols (Wilcoxon signed-rank test, $p = .002$, $n = 10$).¹²
7
8
9
10
11

12 The relationship between peer beliefs about replication success and replication outcomes
13 (i.e., having a “significant” replication) are shown in Figure 3, for prediction market beliefs
14 (Panel A) and survey beliefs (Panel B). Both the prediction market beliefs ($r = .07$, $p = .780$, $n =$
15 20) and the survey beliefs ($r = -.14$, $p = .544$, $n = 20$) were weakly and negatively correlated with
16 replication outcomes. The prediction market and survey beliefs were strongly and positively
17 correlated ($r = .677$, $p = .001$, $n = 20$). Note that these correlation results are based on
18 interpreting the 20 survey and prediction market predictions as independent observations, which
19 may not hold as the predictions may be correlated within the two sets of protocols of each study.
20 Pooling beliefs across protocols so that we have just 10 observations elicits a point-biserial
21 correlation of $-.02$ ($p = .956$) between the prediction market beliefs and replication outcomes, -
22 $.09$ ($p = .812$) between the survey beliefs and the replication outcomes, and $.707$ ($p = .022$)
23 between the prediction market beliefs and the survey beliefs.
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53

54
55 ¹² This survey question was phrased in the following way: “How likely do you think it is that this
56 hypothesis will be replicated (on a scale from 0% to 100%)?”
57
58
59
60

Figure 3 - Peer beliefs about replication outcomes



Note: studies in panel A are ordered according to the prediction market prices for the revised versions. That order is preserved in panel B.

Discussion

We tested whether revising protocols based on formal peer review by experts could improve replication success for a sample of studies that had mostly failed to replicate in a previous replication project (Open Science Collaboration, 2015). Across 10 sets of replications and 13,955 participants from 59 data collection sites, we found that Revised protocols elicited very similar effect sizes as the replication protocols from RP:P. Neither of our primary analysis strategies rejected the null hypothesis that formal peer review has no effect on replicability, and the estimated effect sizes were very small with very narrow confidence intervals ($\Delta r = .002$, 95% CI [-0.04, .04]; $\Delta r = .014$, 95% CI [-0.02, .05]). Both the Revised and the RP:P replications

1
2
3 provided evidence for statistical inconsistency with the original study even across the varied
4 contexts in which multiple labs conducted their replications (Mathur & VanderWeele, 2020).
5
6

7 Ignoring the formal analyses, there was an interesting heuristic pattern that might appear
8 to suggest that formal peer review could improve replicability. Two of the revised protocols
9 showed statistically significant results ($p < .05$) whereas none of the RP:P protocols showed
10 statistically significant results. By comparison, from the exploratory analyses, based on the
11 original effect size and new samples, the average expected percentages of significant results
12 among Revised and RP:P replications were 90% and 92% (i.e., 9 of 10 replications), respectively
13 (Mathur & VanderWeele, 2020). However, even focusing on the significance of these two
14 findings does not provide good evidence for peer review strengthening replication effect sizes.
15 Just 1 of the 2 showed significant moderation by protocol version and, for these two findings, the
16 observed effect sizes for the Revised protocols were an average of 72% smaller than the original
17 findings.
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32

33 Finally, considering the cumulative evidence of the original, RP:P, and the present data,
34 four of the findings had significant effects in the same direction as the original finding, albeit
35 with very small effect sizes. None exceeded $r = .15$ even though the original effect sizes had a
36 median of .37 and a range of .19 to .50. All were quite precisely estimated with the upper bound
37 of the 95% confidence intervals being .23 or less. Considering the 111 effect sizes of all
38 replication sites from RP:P and this investigation, only 4 of them were as large or larger than the
39 original effect size of the finding that they were replicating (see Figure 3). Indeed, of Revised
40 and RP:P replications, exploratory analyses suggested that 50% and 80% respectively provided
41 fairly strong evidence for inconsistency with the original study ($P_{\text{orig}} < .05$), and 20% and 30%
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

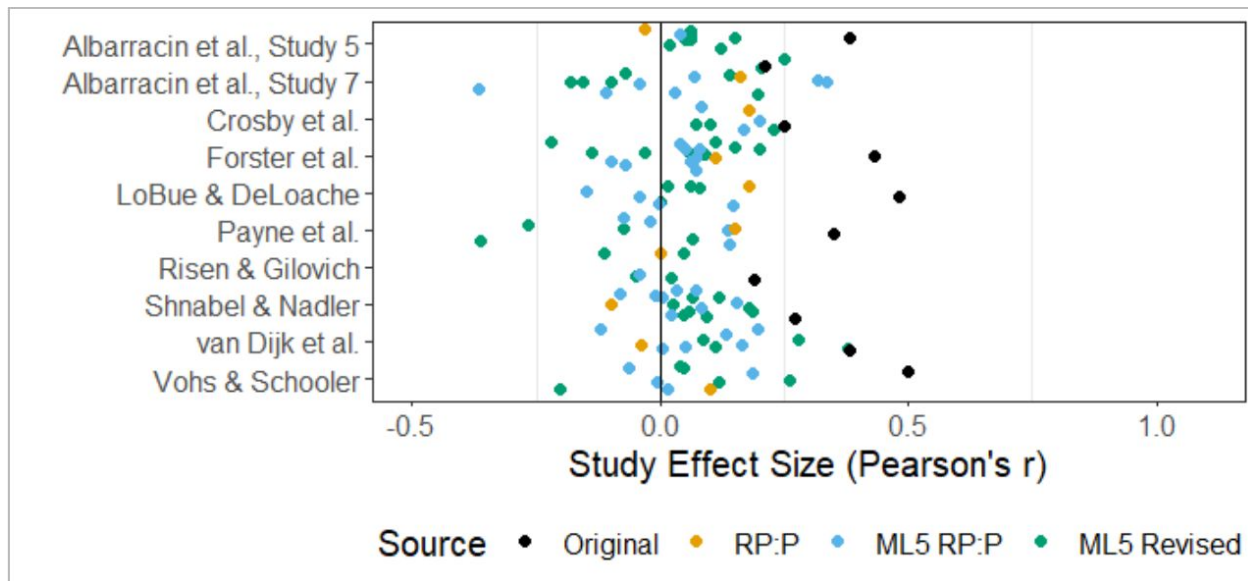
1
2
3 respectively provided strong evidence for inconsistency ($P_{\text{orig}} < .01$). In sum, original effect sizes
4
5 were extreme compared to all attempts to reproduce them.
6

7
8 Conducting formal peer review did not increase observed effect sizes for replication
9
10 efforts of original findings, on average. For a few studies, we observed some evidence consistent
11
12 with the original findings, but with sharply lower effect sizes no matter which protocol was
13
14 considered. This suggests that factors other than expertise that can be communicated through
15
16 peer review are responsible for the substantial difference in observed effect sizes between these
17
18 10 original findings and replication efforts.
19
20

21
22 Finally, neither prediction markets nor surveys performed well in predicting the
23
24 replication outcomes and peer beliefs were not correlated with replication outcomes. Previous
25
26 projects measuring peer beliefs with similar methods have been more successful in predicting
27
28 replication outcomes. One reason for the lower success here could potentially be the small
29
30 sample size of traders and studies producing uncertain estimates (past markets have involved 40-
31
32 80 traders and 20-30 studies, Dreber et al., 2015; Forsell et al., 2018). Also, a floor effect may be
33
34 occurring in that the replications were all much smaller than the original studies providing little
35
36 variability for successful prediction.
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Figure 3 - Effect sizes from individual sites across original studies, RP:P (2015), and Many Labs

5



Specific Implications for Replicability of These 10 Findings

27
28
29
30
31
32
33
34
35
36
37
38

Gilbert et al. (2016) suggested that if the RP:P replication teams had effectively addressed experts' concerns about the designs for these studies and had conducted higher powered tests, then they would have observed replicable results. The present evidence provides mixed support at best for Gilbert et al.'s speculation.

39
40
41
42
43
44
45
46
47
48
49

The most optimistic take would focus on vote counting on achieving statistical significance at $p < .05$. From that perspective, the replication rate went from 0 of these 10 with the RP:P protocol to 2 of 10 with the Revised protocol. Descriptively, it is easy for the optimist to conclude that adding peer review in advance and increasing power substantially increased replicability of the findings.

50
51
52
53
54
55
56
57
58
59
60

The most pessimistic take would counter that even with extremely high power, the formal analyses did not find support that peer review increased replicability across studies. Even focusing on the significant results, only one of the two had evidence consistent with that

1
2
3 hypothesis. Moreover, 3 of the 10 Revised protocols had effects in the opposite direction of the
4 original finding, despite high power and peer review. And, perhaps most critically, effect sizes
5 were dramatically smaller in these optimized replications compared to the original findings. The
6 median effect size for original findings was .37, for RP:P was .11, for new RP:P was .04, and for
7 Revised protocols was .05. On average, original studies would have had 22% power to detect the
8 effect sizes produced by the corresponding Revised protocols (excluding Revised protocols that
9 produced negative effect sizes). Descriptively, it is easy for the pessimist to conclude that adding
10 power and peer review did not help very much, if at all.
11
12
13
14
15
16
17
18
19
20
21

22 The reality is probably somewhere in between the optimistic and pessimistic conclusions.
23
24 The middle of the road perspective might focus on the cumulative evidence. We added a
25 substantial amount of data to the evidence about each of these findings. Figure 2 shows that, with
26 all data combined, 4 of 10 have statistically significant effects ($p < .05$), and all 10 effect sizes
27 are quite precisely estimated and small (median $r = .07$; range 0 to .15). All 10 of the meta-
28 analytic results are much smaller than the original findings (median $r = .37$; range .19 to .50). As
29 data are accumulated, reliable results should be associated with p -values approaching zero rather
30 than remaining close to .05 indicating weak evidence (Benjamin et al., 2017). However, even
31 when retaining the original study, the 4 significant meta-analytic results do not have uniformly
32 very small p -values approaching zero (Crosby et al., 2008, $p = .0004$; Shnabel & Nadler, 2008, p
33 = .015; Albarracin et al., 2008, Study 5, $p = .014$; van Dijk et al., 2008, $p = .023$). The most
34 encouraging individual finding for demonstrating replicability is Crosby et al. (2008). None of
35 the replication studies achieved statistical significance on their own, but the cumulative evidence
36 supports the original finding, albeit with a reduced effect size. Notably, this finding
37 simultaneously showed no evidence of improved replicability based on peer review (the revised
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 protocol elicited an effect size 44% weaker than the original study). The most parsimonious
4
5 explanation for the observed data may be that finding is weaker than indicated by the original
6
7 study and not moderated by the factors that differ between the protocols.
8
9

10 In summary, some of the findings may be replicable and all effect sizes appear to be very
11
12 small, even across the varied contexts in which labs conducted their replications. It is quite
13
14 possible that future replications and refinements of the methodologies supporting these findings
15
16 will yield more significant findings and larger effect sizes. The study that provided the strongest
17
18 evidence for improvement through expert review (Shnabel & Nadler, 2008) provides a
19
20 suggestive direction for such refinements. The primary revisions to that study involved extensive
21
22 tailoring and piloting of study materials for new populations. However, this was not the only
23
24 study whose revisions included this process, reinforcing the possibility that the apparent benefits
25
26 for this finding occurred by chance. Across all studies, the original findings were statistically
27
28 anomalous compared with all replication findings, and the prediction markets, reviewers, and
29
30 replication teams could not predict which findings would persist with some supporting evidence.
31
32
33
34

35 For those findings that failed to improve in replicability, the present understanding of the
36
37 conditions needed for replicating the effect is not sufficient. This minimally suggests that
38
39 theoretical revisions are needed in understanding the boundary conditions for observing the
40
41 effect, and maximally suggests that the original result was a false positive. In the latter case, it is
42
43 possible that no amount of expertise could have produced a replicable finding. We cannot
44
45 definitively parse between these possibilities, but the fact that even protocols revised with formal
46
47 peer review from experts failed to replicate the original effects suggests that theoretical
48
49 understanding of the findings is too weak to specify replicable conditions (Nosek & Errington,
50
51
52
53
54 2020).
55
56
57
58
59
60

Constraints on Generality

There are two primary and related constraints on the generality of our conclusions for the role of expertise in peer review beyond our examined findings: the selection of studies investigated and statistical power. The studies investigated in this project were selected because there was reason, *a priori*, to suspect they could be improved through peer review. If the labeling of these studies as “non-endorsed” accurately reflected serious design flaws, that could mean that our estimate of the effect of peer review represents the extreme end of what should be expected. Conversely, a study selection procedure based on perceived non-endorsement from original authors might have selected for less reliable effects, suppressing the estimate of the effectiveness of peer review. Ultimately, the studies were not selected to be representative of any particular population. The extent to which these findings will generalize is unknown. It is possible that these findings are unique to this sample of studies, or to just psychology studies that are conducted in good faith but fail to be endorsed by original authors as in RP:P (OSC, 2015). A more expansive possibility is that the findings will be generalizable to occasions in which original authors or other experts dismiss a failed replication for having design flaws that are then addressed and tested again. Ultimately, we expect that the findings are partially generalizable in that some expert-guided revisions to research designs will not result in improved replicability. And, we expect that future research will identify boundary conditions on this finding in that some expert-guided revisions to research designs will improve replicability under some conditions. It is unknown whether the conditions under which one or the other will be observed will ever be predictable in advance.

Similarly, the statistical power of the current project limits confidence in the generality of the results. Our study selection criteria and available resources limited us to 10 sets of

1
2
3 replications. Despite our large overall sample size, the number of effect size estimates ($k = 101$)
4
5 and studies investigated (10) might not have afforded sufficient opportunity of diverse conditions
6
7 to observe an effect of peer review. As such, the results of this project should be interpreted as an
8
9 initial, but not definitive, estimate of the effect of pre-data collection peer review on replicability.
10
11

12 **Conclusion: Is Expertise Irrelevant?**

13
14 Concluding that expertise is irrelevant for achieving replicable results may be tempting
15
16 given the very small effect of expert peer review we observed on replication effect sizes.
17
18 However, that interpretation is unwarranted. The present study is a narrow but important test of
19
20 the role of expertise in improving replicability. Our control condition was a set of replications
21
22 using protocols that had mostly failed to replicate in a prior replication project, RP:P. Those
23
24 original replication protocols were developed in a structured process with original materials and
25
26 preregistration; replication researchers had sufficient self-identified expertise to design and
27
28 conduct the replications; and, designs received informal review by an internal review process and
29
30 by original authors when they were willing to provide it. This process did not preclude the
31
32 possibility of errors, but using RP:P protocols meant that the control condition included
33
34 substantial effort and expertise to conduct a faithful replication. Whether that effort and expertise
35
36 was sufficient was the open question. The intervention we tested to improve replicability is a
37
38 function of a particular critique of those failures-to-replicate -- that failure to resolve issues
39
40 identified by original authors signaled critically problematic features of the replication designs.
41
42 So, our finding that formal peer review did not systematically improve replicability may be
43
44 limited to circumstances in which there are already good efforts to conduct high quality
45
46 replications, such as the variety of systematic replication efforts populating social-behavioral
47
48 sciences this decade.
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 It may also be tempting to use the present findings to conclude that conducting formal
4 peer review in advance of conducting studies is not useful for improving quality and credibility.
5 That interpretation is also unwarranted. A possible reason that we failed to replicate some of
6 these findings in presumably ideal circumstances is that the original findings were false
7 positives. If so, then this study does not offer a test of the effectiveness of peer review to improve
8 the quality of study methodology. A finding must be replicable under some conditions to test
9 whether different interventions are influential on its replicability. For several findings, we did not
10 observe any conditions under which these studies were replicable (see also Klein et al., 2019).
11
12
13
14
15
16
17
18
19
20
21

22 There may be conditions under which these studies are more replicable, but peer review
23 did not produce them. Peer reviewers were selected for their perceived expertise in the areas of
24 study we investigated. In many cases, the reviewers authored the original research. It is possible,
25 despite the presumed expertise of the reviewers, that they lacked knowledge of what would make
26 the studies replicable. Other experts may have advised us differently and produced protocols that
27 improved replicability. The current investigation cannot rule out this possibility.
28
29
30
31
32
33
34

35 Finally, it is obvious that expertise matters under a variety of conditions and that lack of
36 expertise can have deleterious impacts in specific cases. For example, conducting an eye-
37 tracking study (e.g., Crosby et al., 2008) minimally requires possessing eye-tracking equipment
38 and having sufficient experience with the equipment to operate it properly. Further, replications
39 can fail for technical reasons; experts may be better positioned to identify those technical errors
40 based on experience with instrumentation and protocols. The meaningful question of the role of
41 expertise for replicability is in the zone that replication researchers appear to possess the basic
42 facility for conducting research of that type, and when those replication researchers perceive that
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 they are conducting an effective replication in good faith. That was the circumstance studied in
4
5 this investigation, and this investigation is hardly the final word.
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Box: Case studies for generating hypotheses about expertise

In the aggregate, our findings indicated little impact of expert peer review on improving replicability across the 10 findings we examined. Nevertheless, a look at individual studies provides occasion for generating hypotheses that could be examined systematically in the future (see Table 1 for descriptions of the difference between protocols).

<p><u>van Dijk et al., 2008</u> van Dijk and colleagues (2008) observed that individuals made more generous offers in negotiations with happy negotiation partners compared to angry negotiation partners ($r = .38$). The Revised protocol ($r = .23$) seemed to elicit an effect more consistent with the original study than did the RP:P protocol did ($r = .06$), but the difference between protocols was not significant ($p = .315$). However, if the difference between protocols for the van Dijk et al (2008) finding is itself replicable, then this paradigm might provide a useful context for investigating the role of expertise systematically. A prior effort to systematically investigate the role of expertise left the question untested because there was little evidence for the phenomenon whether experts guided the protocol development or not (Klein et al., 2019).</p>	<p><u>Payne et al., 2008</u> Payne et al. (2008) observed that implicit and explicit race attitudes were less strongly correlated when participants were told to respond without bias compared to when they were told to express their true feelings ($r = .35$). Replications of this study provide the most curious pattern of all. The original RP:P replication did elicit a significant, but smaller effect than the original study ($r = .15$), but the higher-powered replications with the RP:P ($r = .05$) and Revised ($r = -.16$) protocols did not. In fact, the Revised protocol effect size was in the wrong direction and was significantly different from the RP:P protocol ($p = .002$). Most provocatively, this pattern is directly opposing our hypothesis that formal peer review can improve replicability. More likely, we suspect, is that the finding is weaker than originally observed, non-existent (the original was a false positive), or that the social context for observing the finding has changed.</p>
<p><u>Shnabel & Nadler, 2008</u> Shnabel and Nadler (2008) observed that individuals expressed more willingness to reconcile after a conflict if their psychological needs were restored ($r = .27$). The RP:P ($r = .02$) and Revised ($r = .09$) protocols both elicited substantially weaker effect sizes, but the Revised protocol was slightly larger than the RP:P protocol ($p = .012$). On its own, this pattern is most consistent of all 10 studies with the hypothesis that expert review improves replicability. Even so, the results from Revised replications were not entirely consistent with the original study and yielded a point estimate that was 67% smaller. That just one of the 10 studies showed this effect does increase the plausibility that this one</p>	<p><u>Albarracin et al. 2008</u> Two of our included studies came from Albarracin et al. (2008) that reported evidence that instilling action or inaction goals influences subsequent motor and cognitive output (Study 5 $r = .38$; Study 7 $r = .21$). In RP:P, both studies failed to replicate on the statistical significance criterion, but Study 7's effect size ($r = .16$) was close to the original. Study 5's replication elicited a small effect size in the opposite direction ($r = -.03$). The present replications likewise elicited small effect sizes, but with an interesting pattern. For Study 5, expert review was descriptively and not significantly ($p = .601$) associated with a larger effect size (RP:P $r = .04$; Revised $r = .09$). For Study 7, expert</p>

1
2
3
4 occurred by chance. Nevertheless, if the
5 difference is replicable, then these protocols
6 might help study the role of manipulation
7 checks and effective implementation of the
8 experimental intervention. In this case, the
9 manipulation checks for both protocols
10 suggested that the intervention was effective
11 (Baranski et al., this issue) and yet the
12 outcomes on the dependent variable landed on
13 opposing sides of the statistical significance
14 criterion (p 's = .004, .350).
15
16
17

review was descriptively and not significantly
($p = .150$) associated with an effect size in the
wrong direction (RP:P $r = .02$; Revised $r = -$
.07). If these patterns are not just statistical
noise, it would generate an occasion for
pursuing a perspectivist approach for
understanding the role of expertise in
replicability (McGuire, 2004). Under what
conditions does expertise improve versus
reduce replicability?
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

References

- Albarracín, D., Handley, I. M., Noguchi, K., McCulloch, K. C., Li, H., Leeper, J., ... & Hart, W. P. (2008). Increasing and decreasing motor and cognitive output: a model of general action and inaction goals. *Journal of Personality and Social Psychology*, *95*(3), 510-523.
- Anderson, C. J., Bahník, Š., Barnett-Cowan, M., Bosco, F. A., & Chandler, J. Chartier, CR, ... Zuni, K. (2016). Response to Comment on Estimating the reproducibility of psychological science. *Science*, *351* (6277), 1037-1039. Aad9163.
- Andrews, I., & Kasy, M. (2019). Identification of and correction for publication bias. *American Economic Review*, *109*(8), 2766-94.
- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.--J., Berk, R., ... & Johnson, V. E. (2017). Redefine Statistical Significance. *Nature Human Behavior*, *2*, 6-10. doi:10.1038/s41562-017-0189-z
- Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., ... & Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication?. *Journal of Experimental Social Psychology*, *50*, 217-224.
- Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., ... & Heikensten, E. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433-1436.
- Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... & Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637-644.
- Chambers, C. D. (2013). Registered reports: a new publishing initiative at Cortex. *Cortex*, *49*(3), 609-610.

1
2
3 Chen, S., Szabelska, A., Chartier, C. R., Kekecs, Z., Lynott, D., Bernabeu, P., ... Oberzaucher,
4
5 E. (2018, November 6). Investigating Object Orientation Effects Across 14 Languages.

6
7 <https://doi.org/10.31234/osf.io/t2pju>

8
9
10 Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., ... & Zhou, X. (2018).

11
12 Estimating the reproducibility of experimental philosophy. *Review of Philosophy and*
13
14 *Psychology*, 1-36.

15
16
17 Crosby, J. R., Monin, B., & Richardson, D. (2008). Where do we look during potentially
18
19 offensive behavior?. *Psychological Science*, 19(3), 226-228.

20
21 Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A.,
22
23 & Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of
24
25 scientific research. *Proceedings of the National Academy of Sciences*, 112(50), 15343–
26
27 15347.

28
29
30 Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B.,

31
32 ... & Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the
33
34 academic semester via replication. *Journal of Experimental Social Psychology*, 67, 68-82.

35
36
37 Fleiss JL, Tytun A, Ury HK (1980): A simple approximation for calculating sample sizes for
38
39 comparing independent proportions. *Biometrics*, 36, 343–346.

40
41
42 Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, D., Chen, Y., Nosek, B.A.,

43
44 Johannesson, M., Dreber, A. (2018). Predicting replication outcomes in the Many Labs
45
46 2 study. *Journal of Economic Psychology*.

47
48
49 Förster, J., Liberman, N., & Kuschel, S. (2008). The effect of global versus local processing

50
51 styles on assimilation versus contrast in social judgment. *Journal of Personality and*
52
53 *Social Psychology*, 94(4), 579.

- 1
2
3 Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(6),
4
5 460.
6
7
8 Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on “Estimating the
9
10 reproducibility of psychological science”. *Science*, *351*(6277), 1037-1037.
11
12 Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological*
13
14 *Bulletin*, *82*(1), 1-20.
15
16
17 Harrell Jr, M. F. E. (2019). Package ‘Hmisc’. [http://cran.r-](http://cran.r-project.org/web/packages/Hmisc/Hmisc.pdf)
18
19 [project.org/web/packages/Hmisc/Hmisc.pdf](http://cran.r-project.org/web/packages/Hmisc/Hmisc.pdf)
20
21
22 John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable
23
24 research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524-
25
26 532.
27
28
29 LoBue, V., & DeLoache, J. S. (2008). Detecting the snake in the grass: Attention to fear-relevant
30
31 stimuli by adults and young children. *Psychological Science*, *19*(3), 284-289.
32
33
34 Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C. A., Nosek, B. A., Chartier, C. R., ...
35
36 Ratliff, K. A. (2019, December 11). Many Labs 4: Failure to Replicate Mortality Salience
37
38 Effect With and Without Original Author Involvement.
39
40 <https://doi.org/10.31234/osf.io/vef2c>
41
42
43 Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... &
44
45 Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology*, *45*, 142-
46
47 152.
48
49
50 Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., ... & Nosek,
51
52 B. A. (2018). Many labs 2: Investigating variation in replicability across sample and
53
54 setting. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443-490.
55
56
57
58
59
60

- 1
2
3 Luttrell, A., Petty, R. E., & Xu, M. (2017). Replicating and fixing failed replications: The case of
4
5 need for cognition and argument quality. *Journal of Experimental Social Psychology*, *69*,
6
7 178-183.
8
9
- 10 Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How
11
12 often do they really occur?. *Perspectives on Psychological Science*, *7*(6), 537-542.
13
14
- 15 Mathur, M. B. & VanderWeele, T. J. (2020). New statistical metrics for multisite replication
16
17 projects. *Journal of the Royal Statistical Society: Series A*.
18
19 <https://doi.org/10.1111/rssa.12572>
20
21
- 22 McGuire, W. J. (2004). A perspectivist approach to theory construction. *Personality and Social*
23
24 *Psychology Review*, *8*, 173-182.
25
- 26 Murray, S. L., Derrick, J. L., Leder, S., & Holmes, J. G. (2008). Balancing connectedness and
27
28 self-protection goals in close relationships: A levels-of-processing perspective on risk
29
30 regulation. *Journal of Personality and Social Psychology*, *94*(3), 429-459.
31
32
- 33 Noah, T., Schul, Y., & Mayo, R. (2018). When both the original study and its failed replication
34
35 are correct: Feeling observed eliminates the facial-feedback effect. *Journal of Personality*
36
37 *and Social Psychology*, *114*(5), 657-664.
38
39
- 40 Nosek, B. A., & Errington, T. M. (2017). Reproducibility in cancer biology: making sense of
41
42 replications. *Elife*, *6*, e23383.
43
44
- 45 Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLOS Biology*, *18*, e3000691.
46
47 Doi: 10.1371/journal.pbio.3000691.
48
- 49 Nosek, B. A., & Gilbert, E. A. (2016). Let's not mischaracterize the replication studies.
50
51 *Retraction Watch*, *9*.
52
53
54
55
56
57
58
59
60

1
2
3 Nosek, B. A. & Lakens, D. (2014) Registered Reports: A method to increase the credibility of
4
5 published results. *Social Psychology*, 45, 137-141.
6

7
8 Open Science Collaboration. (2015). Estimating the reproducibility of psychological science.
9
10 *Science*, 349(6251), aac4716.
11

12 Patil, P., Peng, R. D., & Leek, J. T. (2016). What should researchers expect when they replicate
13
14 studies? A statistical view of replicability in psychological science. *Perspectives on*
15
16 *Psychological Science*, 11(4), 539-544
17
18

19 Payne, B. K., Burkley, M. A., & Stokes, M. B. (2008). Why do implicit and explicit attitude tests
20
21 diverge? The role of structural fit. *Journal of Personality and Social Psychology*, 94(1),
22
23 16-31.
24
25

26 Petty, R. E., & Cacioppo, J. T. (2016). Methodological choices have predictable consequences in
27
28 replicating studies on motivation to think: Commentary on Ebersole et al. (2016). *Journal*
29
30 *of Experimental Social Psychology*, 67, 86-87.
31
32

33 Risen, J. L., & Gilovich, T. (2008). Why people are reluctant to tempt fate. *Journal of*
34
35 *Personality and Social Psychology*, 95(2), 293-307.
36
37

38 Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological*
39
40 *Bulletin*, 86(3), 638-641.
41

42 Schwarz, N., & Strack, F. (2014). Does merely going through the same moves make for a
43
44 “direct” replication? Concepts, contexts, and operationalizations. *Social Psychology*,
45
46 45(4), 305-306.
47
48

49 Shnabel, N., & Nadler, A. (2008). A needs-based model of reconciliation: satisfying the
50
51 differential emotional needs of victim and perpetrator as a key to promoting
52
53 reconciliation. *Journal of Personality and Social Psychology*, 94(1), 116-132.
54
55
56
57
58
59
60

- 1
2
3 Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed
4 flexibility in data collection and analysis allows presenting anything as significant.
5
6 *Psychological Science*, 22(11), 1359-1366.
7
8
9
10 Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed
11 addition to all empirical papers. *Perspectives on Psychological Science*, 12(6), 1123-
12 1128.
13
14
15
16
17 Skorb, L., Aczel, B., Bakos, B., Christ, O., Fedor, A., Feinberg, L., Halasa, E., Jiménez-Leal, W.,
18
19 Kauff, M., Kovacs, M., Krasuska, K. K., Kuchno, K., Manfredi, D., Muda, R., Nave, G.,
20
21 Pėkala, E., Pieńkosz, D., Ravid, J., Rentzsch, Katrin, Salamon, J., Schultze, T., Sioma,
22
23 B., & Hartshorne, J. K. (provisionally accepted). Many Labs 5: Replication Report for
24
25 Van Dijk, Van Kleef, Steinel, & Van Beest (2008). A social functional approach to
26
27 emotions in bargaining: When communicating anger pays and when it backfires. *Journal*
28
29 *of Personality and Social Psychology*, 94(4), 600-614. *Advances in Methods and*
30
31 *Practices in Psychological Science*.
32
33
34
35 Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal
36
37 context on picture recognition. *Psychological Science*, 12(2), 153-156.
38
39
40 Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from
41
42 tests of significance—or vice versa. *Journal of the American Statistical Association*,
43
44 54(285), 30-34.
45
46
47 Strack, F. (2016). Reflection on the smiling registered replication report. *Perspectives on*
48
49 *Psychological Science*, 11(6), 929-930.
50
51
52 Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication.
53
54 *Perspectives on Psychological Science*, 9(1), 59-71.
55
56
57
58
59
60

1
2
3 Van Dijk, E., Van Kleef, G. A., Steinel, W., & Van Beest, I. (2008). A social functional
4 approach to emotions in bargaining: When communicating anger pays and when it
5 backfires. *Journal of Personality and Social Psychology*, 94(4), 600-614.
6
7
8
9

10 Vohs, K. D., & Schooler, J. W. (2008). The value of believing in free will: Encouraging a belief
11 in determinism increases cheating. *Psychological Science*, 19(1), 49-54.
12
13
14

15 Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream.
16 *Behavioral and Brain Sciences*, 41. <https://doi.org/10.1017/S0140525X170019>
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table 1 - Summary of main protocol differences

Study	Preregistration	Main Differences between RP:P and Revised Protocol
Albarracín et al., Study 5	osf.io/a3pwa/	RP:P protocol collected participants online from MTurk; Revised protocol collected participants at universities.
Albarracín et al., Study 7	osf.io/725ek/	Original authors expressed concern about replicating the study among participants; original materials were validated in English. Both protocols used only English language participants. Additionally, Revised protocol used scrambled sentences to prime target words because word fragments did not often elicit target words in RP:P replication. RP:P protocol used word fragments.
Crosby et al.	osf.io/tj6qh/	Original authors were concerned that participants in RP:P protocol would be unresponsive to scenarios (concerning affirmative action). Revised protocol presented participatory scenarios after they watched a video about affirmative action. RP:P protocol did not include affirmative action.
Forster et al.	osf.io/ev4nv/	The RP:P replication failed at achieving target ambiguity and applicability of stimuli. In the Revised protocol, stimuli were piloted for both aspects at all collection sites; the RP:P protocol was piloted as the previous RP:P replication.
LoBue & DeLoache	osf.io/68za8/	Original authors expressed concerns regarding the physical features of the control stimuli in the RP:P replication, the age of children recruited, and technical issues such as screen size and Internet speed. The Revised protocol used frogs as control stimuli; the RP:P protocol used control stimuli. In addition, the Revised protocol sampled only 3-year-olds along with a group of 3-5-year-olds, as in the RP:P protocol). Finally, the study was implemented with a different software (allowing the study to be run offline and therefore not be hampered by a smaller screen, more similar to those used in the original studies).
Payne et al.	osf.io/4f5zp/	RP:P protocol collected at sites in Italy in Italian; Revised protocol collected at sites in English.
Risen & Gilovich	osf.io/xxf2c/	RP:P protocol recruited subjects on Amazon Mechanical Turk (MTurk) instead of universities as in original study. Authors of original study were concerned that MTurk experimental scenarios less personally salient than original sample and may be more easily distracted, compromising the cognitive load manipulation. Revised protocol used university students.
Shnabel & Nadler	osf.io/q85az/	In the RP:P protocol, participants read a vignette describing an employee who told their partner to go on a honeymoon; in the Revised protocol, participants read a vignette describing a college student who, upon returning from a two-week family visit, was told by their parents that they found someone who could commit to paying next year's rent and that the protagonist would be the end of the lease. This revision was meant to provide a more relatable experience for participants. The perpetrator of a transgression. The revised materials were created through a pilot study with university students.
van Dijk et al.	osf.io/xy4ga/	Following the original study, the Revised protocol excluded subjects who had taken economics courses or participated in prior psychology studies. Participants were not allowed to see or hear one another during the experiment. These restrictions were not present in the RP:P protocol.

1
2 Vohs & Schooler

osf.io/peuch/

The Revised protocol used different free-will-belief inductions (a rewriting task with text from both pulled from the same source) and a revised measure of free-will-belief (a new instrument) than the RP:P protocol.

3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1 *Table 2 - Summary of sample sizes and power across studies*

2

3	Original Study			RP:P Replication			ML5: RP:P Protocol			ML5: Revised I				
	<i>N</i>	Power to detect ML5 RP:P	Power to detect ML5 Revised	<i>N</i>	Power to detect ML5 RP:P	Power to detect ML5 Revised	Number of Sites	Total <i>LN</i>	Power to detect original ES	Smallest ES with 90% Power	Number of Sites	Total <i>LN</i>	Power to detect original ES	
4														
5														
6														
7														
8	Study													
9	Albarracin et al., Study	36	0.06	0.08	88	0.07	0.13	1	580	> 0.99	0.13	8	884	> 0
10	Albarracin et al., Study				10									
11		98	0.05	0.00	5	0.05	0.00	7	878	> 0.99	0.12	7	808	> 0
12														
13	Crosby et al.	25	0.39	0.34	30	0.46	0.40	3	140	> 0.99	0.11	3	136	> 0
14	Forster et al.	82	0.06	0.07	71	0.05	0.06	8	736	> 0.99	0.13	8	720	> 0
15	LoBue & DeLoache	48	0.05	0.06	48	0.05	0.06	4	286	> 0.99	0.19	4	259	> 0
16					18									
17	Payne et al.	70	0.07	0.00	0	0.10	0.00	4	545	> 0.99	0.14	4	558	> 0
18		12			22									
19	Risen & Gilovich	2	0.00	0.00	6	0.00	0.00	1	2811	> 0.99	0.06	4	701	> 0
20					14									
21	Shnabel & Nadler	94	0.06	0.27	1	0.06	0.40	8	1361	> 0.99	0.05	8	1376	> 0
22		10												
23	van Dijk et al.	3	0.09	0.66	83	0.08	0.56	6	436	> 0.99	0.15	4	119	0.
24	Vohs & Schooler	30	0.06	0.06	58	0.06	0.07	4	279	> 0.99	0.19	5	342	> 0
25														

26
27 *Note: power calculations used $\alpha =$*

28 0.05

29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table 3 - Summary of effect sizes across studies

Study	Original Study			RP:P Replication			ML5: RP:P Protocol			ML5: Revised Protocol		
	<i>N</i>	<i>r</i>	<i>95% CI</i>	<i>N</i>	<i>r</i>	<i>95% CI</i>	<i>N</i>	<i>r</i>	<i>95% CI</i>	<i>N</i>	<i>r</i>	<i>95% CI</i>
Albarracin et al., Study 5	36	0.3	.05, .64	88	0.03	-.24, .18	580	0.04	-.04, .12	884	0.09	.03, .14
Albarracin et al., Study 7	98	0.2	.01, .39	10	-	-.03, .34	878	0.01	-.19, .21	808	0.07	-.17, .03
Crosby et al. 10	25	0.2	.02, .46	5	0.16	-.03, .40	140	0.15	-.01, .30	136	0.14	-.08, .34
Forster et al. 13	82	0.4	.23, .59	30	0.18	-.13, .34	736	0.03	-.02, .09	720	0.05	-.07, .16
LoBue & DeLoache 15	48	0.4	.22, .68	71	0.11	-.12, .45	286	0.01	-.19, .21	259	0.04	-.02, .16
Payne et al. 17	70	0.3	.12, .54	48	0.18	.00, .29	545	0.05	-.13, .22	558	0.16	-.44, .13
Risen & Gilovich 20	12	0.1	.01, .36	0	0.15	-.13, .13	281	-	-.08, -	701	0.01	-.13, .11
Shnabel & Nadler 22	94	0.2	.07, .45	14	-	-.27, .07	136			137		
van Dijk et al. 24	10	0.3	.20, .54	1	0.10	-.26, .18	436	0.06	-.06, .18	6	0.09	.04, .14
Vohs & Schooler 26	3	0.5	.15, .74	83	0.04	-.17, .35	279	0.04	-.14, .22	119	0.23	-.01, .44
	30	0	.15, .74	58	0.10	.35	279	0.04	-.14, .22	342	0.05	-.16, .23

Table 4 - Metrics of replication success by study and protocol version.

Study	Subset	k	Estimate (r)	p-value	τ	P_{org}	Probability significance agreement	Percent above 0	Percent above 0.1	Percent above 0.2
Albarracin et al. Study 5	all	9	0.07 [0.01, 0.12]	0.0023	0	0.06	0.98	100	0	0
Albarracin et al. Study 5	RP:P	1	0.04 [-0.04, 0.12]	0.34	0	0.05	0.96	100	0	0
Albarracin et al. Study 5	Revised	8	0.09 [0.03, 0.14]	0.006	0	0.08	0.98	100	0	0
Albarracin et al. Study 7	all	14	-0.02 [-0.11, 0.07]	0.65	0.10	0.12	0.77	50 [0, 71]	21 [0, 64]	0
Albarracin et al. Study 7	RP:P	7	0.01 [-0.16, 0.18]	0.87	0.13	0.25	0.65	57 [0, 86]	29 [0, 57]	14 [0, 81]
Albarracin et al. Study 7	Revised	7	-0.06 [-0.17, 0.05]	0.19	0.06	0.03	0.84	14 [0, 86]	0	0
Crosby et al.	all	6	0.14 [0.07, 0.21]	0.004	0	0.62	0.82	100	100	0
Crosby et al.	RP:P	3	0.15 [-0.01, 0.30]	0.06	0	0.64	0.80	100	100	0
Crosby et al.	Revised	3	0.14 [-0.09, 0.35]	0.12	0	0.61	0.75	100	100	0
Forster et al.	all	16	0.04 [-0.01, 0.09]	0.10	0	<0.001	1	100	0	0
Forster et al.	RP:P	8	0.03 [-0.02, 0.08]	0.18	0	<0.001	1	100	0	0
Forster et al.	Revised	8	0.04 [-0.06, 0.15]	0.36	0.07	0.004	0.99	75 [0, 100]	12 [0, 100]	0
LoBue & DeLoache	all	8	0.02 [-0.06, 0.11]	0.50	0	0.001	1	100	0	0
LoBue & DeLoache	RP:P	4	0.01 [-0.22, 0.24]	0.89	0.05	0.003	0.99	50 [0, 100]	0	0
LoBue & DeLoache	Revised	4	0.04 [-0.03, 0.10]	0.16	0	0.001	1	100	0	0
Payne et al.	all	8	-0.06 [-0.21, 0.09]	0.40	0.16	0.04	0.82	38 [0, 62]	25 [0, 50]	0

1				0.05 [-0.13,					75 [0,		
2	Payne et al.	RP:P	4	0.22]	0.46	0.07	0.03	0.94	100]	50 [0, 100]	0
3											
4											
5	Payne et al.	Revised	4	-0.16 [-0.44,	0.20	0.18	0.03	0.72	25 [0, 50]	0	0
6				0.15]							
7											
8	Risen &										
9	Gilovich	all	5	-0.04 [-0.14,	0.20	0	0.02	0.96	0	0	0
10				0.07]							
11	Risen &										
12	Gilovich	RP:P	1	-0.04 [-0.08,	0.02	0	0.02	0.96	0	0	0
13				-0.01]							
14											
15	Risen &										
16	Gilovich	Revised	4	-0.01 [-0.18,	0.87	0.01	0.06	0.83	0	0	0
17				0.16]							
18	Shnabel &										
19	Nadler	all	16	0.05 [0.02,	0.009	0	0.04	0.99	100	0	0
20				0.09]							
21	Shnabel &										
22	Nadler	RP:P	8	0.02 [-0.03,	0.38	0	0.02	0.98	100	0	0
23				0.08]							
24											
25	Shnabel &										
26	Nadler	Revised	8	0.09 [0.03,	0.008	0	0.08	0.98	100	0	0
27				0.14]							
28											
29	van Dijk et al.	all	10	0.10 [-0.01,	0.07	0.01	0.006	1	100	40 [0, 100]	0
30				0.20]							
31											
32	van Dijk et al.	RP:P	6	0.06 [-0.06,	0.24	0	0.002	1	100	0	0
33				0.19]							
34											
35	van Dijk et al.	Revised	4	0.23 [-0.03,	0.07	0	0.19	0.97	100	100	100
36				0.45]							
37											
38	Vohs &										
39	Schooler	all	9	0.04 [-0.06,	0.37	0.06	0.01	0.98	78	22 [0, 100]	0
40				0.15]							
41											
42	Vohs &										
43	Schooler	RP:P	4	0.04 [-0.15,	0.55	0	0.01	0.98	100	0	0
44				0.23]							
45	Vohs &										
46	Schooler	Revised	5	0.05 [-0.16,	0.55	0.11	0.03	0.94	80	20 [0, 100]	0
47				0.25]							
48											
49											

50 *Study*: Name of original study. *Subset*: all replications for the study, replications using the RP:P
51 *protocol*, or replications using the Revised protocol. *k*: number of studies in subset. *Estimate (r)*
52 *and p-value*: meta-analytic estimate in replications on Pearson's *r* scale with 95% confidence
53 *interval (brackets) and p-value*. *τ*: meta-analytic heterogeneity estimate of standard deviation of
54 *effects in replications*. *P_{orig}*: probability that the original study's estimate would be as extreme as
55 *actually observed if the original study were consistent with the replications*. *Probability*
56 *significance agreement*: probability that the meta-analytic estimate in the replications would be
57
58
59
60

1 statistically significant and would agree in direction with that of the original study if the original
2 and the replications were consistent. Percent above 0, 0.1, and 0.2: estimated percentage of
3 population effects stronger than thresholds of $r = 0, 0.1, \text{ and } 0.2$ respectively. Brackets denote
4 95% confidence intervals, which are omitted for the percentage metrics when they could not be
5 estimated.
6

7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60