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

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

Maya B. Mathur, Quantitative Sciences Unit, Stanford University

Erica Baranski, University of Houston

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

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

Christopher R. Chartier, Department of Psychology, Ashland University

Katherine S. Corker, Grand Valley State University

Martin Corley, Psychology, PPLS, University of Edinburgh

Joshua K. Hartshorne, Department of Psychology, Boston College

Hans IJzerman, LIP/PC2S, Université Grenoble Alpes

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

Hugh Rabagliati, Psychology, PPLS, University of Edinburgh

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

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

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

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

Jack D. Arnal, McDaniel College

Holly Arrow, Department of Psychology, University of Oregon

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

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

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

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

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

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

Michał Białek, Department of Economic Psychology, Kozminski University, Poland

Nicholas G. Bloxsom, Department of Psychology, Ashland University

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

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

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

Florian Brühlmann, Department of Psychology, University of Basel

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

Erica Casini, University of Milano - Bicocca, Italy

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

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

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

Oliver Christ, Fernuniversität in Hagen, Germany

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

Heather M. Claypool, Department of Psychology, Miami University

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

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

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

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

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

Benjamin Dering, Psychology, University of Stirling, UK

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

John E. Edlund, Rochester Institute of Technology

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

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

Lily Feinberg, Department of Psychology, Boston College

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

Máire Ford, Department of Psychology, Loyola Marymount University

Michael C. Frank, Department of Psychology, Stanford University

Emily Fryberger, Department of Psychology, Pacific Lutheran University

Alexander Garinther, Department of Psychology, University of Oregon

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

Kayla Gerken, Rose-Hulman Institute of Technology

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

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

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

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

Ewa Hałasa, Maria Curie-Sklodowska University, Poland

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

Rias A. Hilliard, Rose-Hulman Institute of Technology

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

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

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

Michael Inzlicht, Department of Psychology, University of Toronto

Alan Jern, Rose-Hulman Institute of Technology

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

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

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

Mathias Kauff, Medical School Hamburg, Germany

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

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

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

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

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

Vanessa S. Kolb, Department of Psychology, University of Rhode Island

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

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

Karolina Krasuska, Maria Curie-Sklodowska University, Poland

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

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

Katarzyna Kuchno, Maria Curie-Sklodowska University, Poland

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

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

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

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

Hause Lin, Department of Psychology, University of Toronto

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

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

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

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

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

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

Mitchell Metzger, Department of Psychology, Ashland University

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

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

Jeremy K. Miller, Department of Psychology, Willamette University

Andres Montealegre, Universidad de los Andes

Don A. Moore, University of California at Berkeley

Rafał Muda, Maria Curie-Sklodowska University, Poland

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

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

Sarah A. Novak, Department of Psychology, Hofstra University

Christian Nunnally, Rose-Hulman Institute of Technology

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

Anna Palinkas, Eötvös Loránd University

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

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

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

Emilian Pękala, Maria Curie-Sklodowska University, Poland

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

Sebastiaan Pessers, University of Leuven, Belgium

Boban Petrović, Institute of Criminological and Sociological Research, Serbia

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

Damian Pieńkosz, Maria Curie-Sklodowska University, Poland

Emanuele Preti, University of Milano - Bicocca, Italy

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

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

Jonathan Ravid, Department of Psychology, Boston College

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

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

Juliette Richetin, University of Milano-Bicocca, Italy

Sean C. Rife, Murray State University

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

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

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

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

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

Kathleen Schmidt, Department of Psychology, Southern Illinois University Carbondale

Kurt Schuepfer, Department of Psychology, Miami University

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

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

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

Ani Shabazian, Loyola Marymount University, USA

Rachel L. Shubella, Rose-Hulman Institute of Technology

Adam Siegel, Cultivate Labs

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

Barbara Sioma, Maria Curie-Sklodowska University, Poland

Lauren Skorb, Department of Psychology, Boston College

Luana Elayne Cunha de Souza, University of Fortaleza, Brazil

Sara Steegen, University of Leuven, Belgium

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

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

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

Daniel Storage, Department of Psychology, University of Denver

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

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

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

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

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

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

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

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

Ann-Kathrin Torka, Department of Psychology, TU Dortmund University, Germany

Francis Tuerlinckx, University of Leuven, Belgium

Wolf Vanpaemel, University of Leuven, Belgium

Leigh Ann Vaughn, Department of Psychology, Ithaca College

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

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

Maria Vlachou, University of Leuven, Belgium

Ryan J. Walker, Department of Psychology, Miami University

Sophia C. Weissgerber, Universität Kassel, Germany

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

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

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

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

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

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

Mark Zrubka, Eötvös Loránd University, Hungary

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

Running head: MANY LABS 5

Article type: Registered Report

Received: 12/07/2018

Revision accepted: 08/21/2020

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.

A finding might not replicate for several reasons. The initial finding might have been a false positive, reflecting either a "normal" Type I error or one made more likely through selective reporting of positive results and ignoring null results (Greenwald, 1975; Rosenthal, 1979; Sterling, 1959), or by employing flexibility in analytic decisions and reporting (Gelman & Loken, 2014; John et al., 2012; Simmons, Nelson, & Simonsohn, 2011). Alternatively, the theory being tested might be insufficiently developed, such that it cannot anticipate possible moderators inadvertently introduced in the replication study (Simons, Shoda, & Lindsay, 2017). Finally, the replication study might have been a false negative, reflecting either a lack of statistical power or an ineffective or unfaithful methodology that disrupted detecting the true effect. Many prior replication efforts attempted to minimize false negatives by using large samples, obtaining original study materials, and requesting feedback from original authors on study protocols before they were administered. Nevertheless, these design efforts may not have been sufficient to reduce or eliminate false negatives for true effects. For example, in the Reproducibility Project: Psychology (RP:P; Open Science Collaboration, 2015), replication teams sought materials and feedback from original authors to maximize the quality of the 100 replication protocols. In 11 cases, studies were identified as "not endorsed" meaning that original authors had identified potential shortcomings *a priori* that were not addressed in the ultimate design.[2] These shortcomings may have had implications for replication success. Of the 11 studies, only 1 successfully replicated the original finding, albeit much more weakly than the original study. These unresolved issues were cited in a critique of RP:P as a likely explanation for replication

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

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

Unfaithful or Ineffective Methods as a Moderator of Replicability

Replication is attempting to reproduce a previously observed finding with no *a priori* expectation for a different outcome (see Nosek & Errington, 2020; Nosek & Errington, 2017; Zwaan et al., 2018). Nevertheless, a replication may still produce a different outcome for a variety of reasons (Gilbert, King, Pettigrew, & Wilson, 2016; Luttrell, Petty, & Xu, 2017; Noah, Schul, & Mayo, 2018; Open Science Collaboration, 2015; Petty & Cacioppo, 2016; Stroebe & Strack, 2014; Schwarz & Strack, 2014; Strack, 2016). Replicators could fail to implement key features of the methodology that are essential for observing the effect. They could also administer the study to a population for which the finding is not expected to apply. Alternatively, replicators could implement features of the original methodology that are not appropriate for the new context of data collection. For example, in a study for which object familiarity is a key feature, objects familiar to an original sample in Europe might not be similarly familiar to a new sample in Asia. A more appropriate test of the original question might require selecting new objects that have comparable familiarity ratings across populations (e.g., Chen, Chartier, & Szabelska, 2018, replications of Stanfield & Zwaan, 2001). These simultaneous challenges of (a) adhering to the original study, and (b) adapting to the new context, have the important implication that claims over whether or not a particular study is a replication is theory-laden (Nosek & Errington, 2017; 2020). Since exact replication is impossible, claiming "no *a priori* expectation for a different outcome" is an assertion that all of the differences between the original study and the replication are theoretically irrelevant for observing the identified effect.

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

The studies in RP:P that were "not endorsed" by original authors offer a unique opportunity to test this hypothesis. The RP:P protocols were deemed by the replication teams to be replications of the original studies. However, original authors expressed concerns prior to data collection. Thus, if any failed replications can be explained due to poor replication design, these are among the top candidates. Thus, we revised 10 of the 11 "non-endorsed" protocols from RP:P and subjected them to peer review before data collection, a model known as Registered Reports (Chambers, 2013; Nosek & Lakens, 2014; http://cos.io/rr/). Once the protocols were accepted following formal peer review, they were preregistered on OSF (see Table 1). Then, we conducted replications using both the RP:P protocols and the Revised protocols, with multiple laboratories contributing data for one or both protocols. This "many labs" design allowed us to achieve unusually high statistical power, decreasing the probability that any failure to replicate could be due to insufficient power.

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

would not necessarily mean that *all* failures to replicate are due to poor design -- our sample of studies was chosen because they are among the most likely published replications to have faulty designs -- but it would suggest that published replicability rates are overly pessimistic. Note that the replications using the original RP:P protocols serve as a control: If both protocols lead to successful replications, then the failures in RP:P were more likely due to low power or some unexpected difference in the replication teams themselves. In contrast, if most of the replications fail even after expert input, it casts doubt on the "design error" hypothesis, at least for these studies. Rather, such an outcome would increase the likelihood that the original findings were false positives because even formal expert input had no effect on improving replicability.

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

Disclosures

Confirmatory analyses were preregistered on OSF (https://osf.io/nkmc4/). Links to the preregistrations for the individual replications can be found in Table 1. All materials, data, and code are available on the OSF (https://osf.io/7a6rd/). The RP:P Protocols were created from the original RP:P materials that can be found here: https://osf.io/ezcuj/. We report how we

determined our sample size, all data exclusions, all manipulations, and all measures in the study. Data were collected in accordance with the Declaration of Helsinki. The authors acknowledge a conflict-of-interest that Brian Nosek is Executive Director of the non-profit Center for Open Science that has a mission to increase openness, integrity, and reproducibility of research. This project was supported by a grant from the Association for Psychological Science and from Arnold Ventures. In addition, the following authors thank other sources of funding: the French National Research Agency (ANR-15-IDEX-02; IJzerman), the Netherlands Organization for Scientific Research (NWO) (016.145.049; IJzerman), the National Institute on Alcohol Abuse and Alcoholism (F31AA024358; MH Bernstein), the Social Sciences and Humanities Research Council of Canada (149084; Inzlicht), the Economic and Social Research Council (UK, ES/L01064X/1, Rabagliati), John Templeton Foundation (Ebersole and Nosek), Templeton World Charity Foundation (Nosek), and Templeton Religion Trust (Nosek). The authors thank the many original authors and experts who provided extensive feedback throughout the many stages of the project.

Method

Replications

Our selection criteria for studies to replicate consisted of those labeled “not-endorsed” from RP:P (Open Science Collaboration, 2015). For each of the 11 candidate studies, we sought team leads to conduct the new replications and enough research teams to satisfy our sampling plan (see below). We recruited researchers through professional listservs, personal contacts, and collaboration websites (StudySwap, https://osf.io/view/StudySwap/). We were able to satisfy our recruitment goals for 10 of the 11 replications (all except Murray et al., 2008). For each of the 10 studies, we conducted two replications with multiple samples each: one using the RP:P protocol,

and the other using the Revised protocol that was approved following formal peer review. Because RP:P focused on a single statistical result from each original study, both protocols focused on replicating that same result.

Preparation of Protocols and Peer-Review

Teams reconstructed each RP:P protocol using the methods and materials that were shared by the original RP:P replication teams (https://osf.io/ezcuj/). This protocol was the basis for the RP:P protocol condition. Differences between the RP:P Protocol and the replication as described in RP:P reflected practicalities such as lab space, population, climate, and time of year (see other articles, this issue, for details of those replications). Next, teams sought out any correspondence and/or responses written by the original authors concerning the RP:P replications.[3] Teams revised the RP:P protocols to account for concerns expressed in those sources. This revision was the basis for the Revised protocol condition. Then, both the RP:P protocols and the Revised protocols were submitted for peer-review through *Advances in Methods and Practices in Psychological Science* with the agreement of the Editor that only the Revised protocols would be reviewed and revised based on expert feedback. If the original authors were unavailable or unwilling to provide a review, the Editor sought input from other experts. Based on editorial feedback, teams updated their Revised protocols and resubmitted them for additional review until the protocols were given in-principle acceptance.

The peer review process produced a range of requested revisions across replication studies. Some revisions concerned using a participant sampling frame more similar to that of the original study (e.g., some RP:P protocols differed from original studies regarding sampling from the lab vs. MTurk, different countries, different age ranges). Some revisions increased

[3] Correspondence from RP:P (OSC, 2015) was accessed from that project's OSF page (osf.io/ezcuj/).

methodological alignment of the Revised protocol with that of the original study. Other revisions altered the protocol from the original to make it more appropriate for testing the original research question in the replication contexts. Importantly, we were agnostic to which types of changes would be most likely to yield successful replications. We sought to enact all revisions that experts deemed important to make successful replication as likely as possible and that were feasible given available resources. If there were disagreements about the feasibility of a request, the Editor made a final decision (though this was rare).

Upon acceptance, teams preregistered their protocols on OSF and initiated data collection. Table 1 provides links to the preregistered protocols and brief summaries of the main differences between the RP:P and Revised protocols (i.e., the primary changes to the protocol suggested by reviewers/previous correspondence). The reports for each of the 10 studies were submitted for results-blind review so that the Editor and reviewers could examine how confirmatory analyses would be conducted and presented. To ensure that the authors and reviewers could discuss the current study's methods and analysis plan without being biased by the results, the present summary report was drafted and peer-reviewed prior to the two project organizers knowing the results of the majority of the replications (BAN knew none of the results; CRE was directly involved with data collection for two of the sets of replications and was aware of only those results). CRE and BAN had primary responsibility for drafting the paper, and all other authors contributed to revisions. Other authors knew outcomes of 0 or 1 of the sets of replications during the writing process depending on which individual studies they helped conduct. The full reports of each individual replication are reported separately in this issue [CITATIONS TO BE ADDED]. All data, materials, code, and other supplementary information are available at https://osf.io/7a6rd/.

Sampling Plan

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

Eliciting peer beliefs

Predictions about replication success guided the selection and revision of studies to replicate in this project. To assess whether other researchers shared these predictions, we measured peer beliefs about the replications. Following previous efforts (Dreber et al., 2015; Camerer et al., 2016; 2018; Forsell et al., 2018), we invited psychology researchers to predict the

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

replication outcomes for the 10 RP:P protocols and 10 Revised protocols in prediction markets and surveys. Before being allowed to trade in the markets, participants had to rate the probability of the binary measure of successful replication (a statistically significant effect at $p < 0.05$ in the same direction as the original study) for each of the 20 protocols in a survey. In the prediction market, participants traded contracts worth money if the study replicated and worth nothing if the study did not replicate. With some caveats (Manski, 2006), the prices of such contracts can be interpreted as the probabilities that the market assign the studies replicating. For each study, participants could enter the quantity of the contract they wanted to buy (if they believed that the true probability that the study will replicate is higher than the one specified by the current price) or to sell (if they believed that the true probability that the study will replicate is lower than the one identified by the current price). Participants were endowed with points corresponding to money that we provided, and they thus had a monetary incentive to report their true beliefs. For each study, participants were provided with links to the RP:P protocols, the Revised protocols, and to a document summarizing the differences between the two. They were informed that all the replications had a power of at least 80%. The prediction markets were open for two weeks starting from June 21st, 2017, and a total of 31 participants made at least one trade. See the Supplemental Material for more details about the prediction markets and survey.

Power Analyses

The primary test for this study involved comparing the replicability of studies using protocols from RP:P compared to those using protocols revised through expert peer review. We calculated our power to detect such an effect, measured as the effect of protocol within each set of studies ($k = 10$). The results are displayed in Figure 1.[5] In cases of both low ($I^2 = 25\%$) and

[5] See https://osf.io/j5vnh/ for power and figure script.

moderate (I^2 = 50%) heterogeneity, our minimum planned samples should provide adequate power (> 80%) to detect an average effect of protocol as small as r = .05. For greater heterogeneity (I^2 = 75%), our minimum planned samples should provide adequate power to detect an effect of protocol as small as r = .075. Power under all heterogeneity assumptions approaches 100% for effects of r = .10 or larger. As a comparison, the difference between effect sizes reported in the original studies and those reported in RP:P were, on average, Δr = .27.

We also simulated our estimated power for a second analysis strategy, that being meta-analyzing the effect sizes from each protocol within each individual site and testing protocol version as a meta-analytic moderator.[6] These power estimates were slightly lower.

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 Δr = .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 average correlation of 0.58 between peer beliefs and replication outcomes (78% power with twenty markets).

[6] See https://osf.io/dhr3p/ for power simulation script.

Results

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

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

In the present study, 0 of 10 replications using the RP:P protocol yielded a "statistically significant" meta-analytic effect size ($p < .05$), the median effect size was $r = .04$, and the median sample size was $N = 562.5$. Also in the present study, 2 of 10 replications[7] using the Revised protocol yielded statistically significant meta-analytic effect sizes ($p < .05$), the median effect size was $r = .07$, and the median sample size was $N = 629.5$. Gauging replication success based on whether the replications are statistically significant is subject to substantial caveats. For example, depending on the power of the original study and replications, the expected proportion of significant replications can be quite low even when the original is consistent with the replications (Andrews & Kasy, 2019; Patil, Peng, & Leek, 2016). Therefore, as a benchmark to help interpret these metrics regarding statistical significance, we estimated the expected probability that each pooled replication estimate would be "statistically significant" and positive in sign, if in fact the replications were consistent with the original study (Mathur &

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

VanderWeele, 2020). A full summary of aggregated effect sizes and confidence intervals for each data collection appears in Table 3.

The purpose of this investigation was to test whether a protocol resulting from formal peer review would produce stronger evidence for replicability than a protocol that had not received formal peer review. We tested this in two ways. First, we calculated an effect size for each protocol within each data collection site. Each site implementing both the RP:P protocol and the Revised protocol contributed two effect sizes, and each site implementing only one of the two protocols contributed one effect size. We conducted a multilevel random-effects meta-analysis of the $k = 101$ effect sizes[8], with a random intercept of data collection Site (varying from 3 to 9 depending on Study) nested within Study (10 studies). This model converged so we did not alter the model further. Then, we added the protocol version (RP:P vs. Revised), the hypothesized moderator, as a fixed effect. We found that it had a near zero effect, $b = .002$, $SE = .02$, $z = .091$, $p = .928$, 95% CI [-.04, .04]. That is, effect sizes from Revised protocols were, on average, $b = .002$ units on the Pearson's r scale larger than effect sizes from RP:P protocols. Overall, effect sizes had little variance accounted for by the moderator as indexed by Tau = .05 (95% CI [0, .09]) on the Fisher's z scale. There was, however, significant heterogeneity between the effect sizes overall, as indicated by the Q statistic, $Q = 147.07$, $p = .001$, $I^2 = 26.57\%$.

For the second test, we conducted a random-effects meta-analysis on the estimates of the effect of protocol within each replication. We calculated the strength of the effect of protocol on the Pearson's r scale for each of the 10 studies. A meta-analysis of these $k = 10$ estimates

[8] Throughout, we meta-analyzed effect sizes on the Fisher's z scale, but report results transformed back to the Pearson's r scale for interpretability except where otherwise noted.

suggested that these effect sizes were not reliably different from zero, b = .014, SE = .01, t = .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 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. (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 for each of these findings, as was done in the individual reports for this project, indicates that for only one of the two tests was the revised protocol effect size significantly larger (Albarracin et al., Study 5, p = .601; Shnabel & Nadler, p = .012). Therefore, even looking at the most

promising examples of the effect of formal peer review on increasing replicability fails to provide reliable support. It is possible that the expert feedback did reliably improve the Shnabel and Nadler effect size, but given the number of tests, it is also plausible that this difference occurred by chance.

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

Exploratory Analyses - Additional Measures of Replicability

As exploratory analyses, we considered several other measures of replicability that directly assess: (1) statistical consistency between the replications and the original studies; and (2) the strength of evidence provided by the replications for the scientific effect under investigation (Mathur & VanderWeele, 2020). These analyses also account for potential heterogeneity in replications and for the sample sizes tested in both the replications and the original studies. Accounting for these sources of variability avoids potentially misleading conclusions regarding replication success that can arise from metrics that do not account for these sources of variability, such as agreement in statistical significance.

First, an original study can be considered statistically "consistent" with a set of replications if the original study and the replications came from the same distribution of potentially heterogeneous effects – that is, if the original study was not an anomaly (Mathur &

VanderWeele, 2020). We assessed statistical consistency using the metric P_{orig}. Analogous to a *p*-value for the null hypothesis of consistency, this metric characterizes the probability that the original study would have obtained a point estimate at least as extreme as was observed, if in fact the original study were consistent with the replications. P_{orig} thus assesses whether the replications were similar to those of the original study with small values of P_{orig} indicating less similarity and larger values indicating more similarity.

Second, we assessed the strength of evidence provided by the replications for each scientific hypothesis investigated in the original studies (Mathur & VanderWeele, 2020). Specifically, we estimated the percentage of population effects, among the potentially heterogeneous distribution from which the replications are a sample, that agree in direction with the original study. This metric is generous toward the scientific hypothesis by treating all effects in the same direction as the original study, even those of negligible size, as evidence in favor of the hypothesis. More stringently, we also estimated the percentage of population effects that not only agreed in direction with the original, but were also meaningfully strong by two different criteria (i.e., $r > .10$ or $r > .20$). These metrics together assess whether the replications provided standalone evidence for the scientific hypothesis, regardless of the estimate of the original study itself.

For each study, we conducted these analyses for three subsets: 1) all replications, regardless of which protocol they used; 2) replications using the RP:P protocol; and 3) replications using the Revised protocol. Note that the three percentage metrics should be interpreted cautiously for subsets of fewer than 10 replications that also have heterogeneity estimates greater than 0, and we conducted sensitivity analyses excluding four such studies from aggregated statistics (Mathur & VanderWeele, 2020; see Supplement for methodological

details). For replication subsets that had a heterogeneity estimate of exactly 0 or which had only 1 replication, we simply report the percentage as either 100% or 0% depending on whether the single point estimate was above or below the chosen threshold.

Table 4 aggregates these results, showing the mean value of P_{orig} and the mean percentages of effects stronger than r = 0, .1, and .2 respectively. Despite our close standardization of protocols across sites, 40% of replication sets within each of the 3 subsets had heterogeneity estimates greater than 0, highlighting the importance of estimating heterogeneity when assessing replications. Regarding statistical consistency between the originals and the replications, the median values of P_{orig} were .04 and .02 for the Revised replications and the RP:P replications, respectively. That is, there were on average 4% and 2% probabilities that the original studies' estimates would have been at least as extreme as observed if, for each study, the original and replication studies had come from the same distribution. Of Revised and RP:P replications, 50% and 80% respectively provided fairly strong evidence for inconsistency with the original study ($P_{orig} < .05$), and 20% and 30% respectively provided strong evidence for inconsistency ($P_{orig} < .01$). Thus, both the Revised and the RP:P replications often suggested statistical inconsistency with the original study, even after accounting for effect heterogeneity and other sources of statistical variability.[9] However, heuristically, evidence for inconsistency might have been somewhat less pronounced in the Revised than in the RP:P replications.

Regarding evidence strength for the scientific hypotheses, for the Revised replications, on average only 50% of population effects agreed in direction with the original study (as expected if

[9] We performed sensitivity analyses that excluded the replication subsets that had fewer than 10 replications as well as heterogeneity estimates greater than 0. In these analyses, the median values of P_{orig} were .08 and .01 for the Revised replications and the RP:P replications, respectively. Of Revised and RP:P replications, 20% and 86% respectively had $P_{orig} < .05$, and 20% and 29% respectively had $P_{orig} < .01$.

the average effect size were exactly zero), 20% were above a modest effect size of r = .10, and 10% were above r = .20. For the RP:P replications, on average, 60% of effects agreed in direction with the original study, 10% were above r = .10, and 0% were above r = .20. These results suggest that even after accounting for heterogeneity, the large majority of population effects were negligibly small regardless of protocol version.[10] Thus, in both the Revised and the RP:P replications, the population effects did not reliably support the scientific hypotheses even when generously considering all effects that agreed in direction with the original study as providing support; furthermore, only a small minority of population effects in each case were meaningfully strong in size.

Peer Beliefs

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

The survey and prediction markets produce a collective peer estimate of the replication success probability for each replication. The mean predicted probability of a statistically significant replication was .286 (range .124-.591) for the 10 RP:P protocols and .296 (range .065-.608) for the 10 Revised protocols (Wilcoxon signed-rank test, p = .232, n = 10), implying that participants expected about 3 of 10 studies from each protocol to replicate. The mean survey

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

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

belief about replication success was .335 (range .217-.528) for the 10 RP:P protocols and .367 (range .233-.589) for the 10 Revised protocols (Wilcoxon signed-rank test, $p = .002$, $n = 10$).[12]

The relationship between peer beliefs about replication success and replication outcomes (i.e., having a "significant" replication) are shown in Figure 3, for prediction market beliefs (Panel A) and survey beliefs (Panel B). Both the prediction market beliefs ($r = .07$, $p = .780$, $n = 20$) and the survey beliefs ($r = -.14$, $p = .544$, $n = 20$) were weakly and negatively correlated with replication outcomes. The prediction market and survey beliefs were strongly and positively correlated ($r = .677$, $p = .001$, $n = 20$). Note that these correlation results are based on interpreting the 20 survey and prediction market predictions as independent observations, which may not hold as the predictions may be correlated within the two sets of protocols of each study. Pooling beliefs across protocols so that we have just 10 observations elicits a point-biserial correlation of -.02 ($p = .956$) between the prediction market beliefs and replication outcomes, -.09 ($p = .812$) between the survey beliefs and the replication outcomes, and .707 ($p = .022$) between the prediction market beliefs and the survey beliefs.

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

[12] This survey question was phrased in the following way: "How likely do you think it is that this hypothesis will be replicated (on a scale from 0% to 100%)?"

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 (Δr = .002, 95% CI [-.04, .04]; Δr = .014, 95% CI [-.02, .05]). Both the Revised and the RP:P replications provided evidence for statistical inconsistency with the original study even across the varied contexts in which multiple labs conducted their replications (Mathur & VanderWeele, 2020).

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

Finally, considering the cumulative evidence of the original, RP:P, and the present data, four of the findings had significant effects in the same direction as the original finding, albeit with very small effect sizes. None exceeded r = .15 even though the original effect sizes had a median of .37 and a range of .19 to .50. All were quite precisely estimated with the upper bound of the 95% confidence intervals being .23 or less. Considering the 111 effect sizes of all replication sites from RP:P and this investigation, only 4 of them were as large or larger than the original effect size of the finding that they were replicating (see Figure 3). Indeed, of Revised

and RP:P replications, exploratory analyses suggested that 50% and 80% respectively provided fairly strong evidence for inconsistency with the original study ($P_{orig} < .05$), and 20% and 30% respectively provided strong evidence for inconsistency ($P_{orig} < .01$). In sum, original effect sizes were extreme compared to all attempts to reproduce them.

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

Finally, neither prediction markets nor surveys performed well in predicting the replication outcomes and peer beliefs were not correlated with replication outcomes. Previous projects measuring peer beliefs with similar methods have been more successful in predicting replication outcomes. One reason for the lower success here could potentially be the small sample size of traders and studies producing uncertain estimates (past markets have involved 40-80 traders and 20-30 studies, Dreber et al., 2015; Forsell et al., 2018). Also, a floor effect may be occurring in that the replications were all much smaller than the original studies providing little variability for successful prediction.

Specific Implications for Replicability of These 10 Findings

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.

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.

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

The reality is probably somewhere in between the optimistic and pessimistic conclusions. The middle of the road perspective might focus on the cumulative evidence. We added a substantial amount of data to the evidence about each of these findings. Figure 2 shows that, with all data combined, 4 of 10 have statistically significant effects ($p < .05$), and all 10 effect sizes are quite precisely estimated and small (median $r = .07$; range 0 to .15). All 10 of the meta-

analytic results are much smaller than the original findings (median r = .37; range .19 to .50). As data are accumulated, reliable results should be associated with p-values approaching zero rather than remaining close to .05 indicating weak evidence (Benjamin et al., 2017). However, even when retaining the original study, the 4 significant meta-analytic results do not have uniformly very small p-values approaching zero (Crosby et al., 2008, p = .0004; Shnabel & Nadler, 2008, p = .015; Albarracin et al., 2008, Study 5, p = .014; van Dijk et al., 2008, p = .023). The most encouraging individual finding for demonstrating replicability is Crosby et al. (2008). None of the replication studies achieved statistical significance on their own, but the cumulative evidence supports the original finding, albeit with a reduced effect size. Notably, this finding simultaneously showed no evidence of improved replicability based on peer review (the revised protocol elicited an effect size 44% weaker than the original study). The most parsimonious explanation for the observed data may be that finding is weaker than indicated by the original study and not moderated by the factors that differ between the protocols.

In summary, some of the findings may be replicable and all effect sizes appear to be very small, even across the varied contexts in which labs conducted their replications. It is quite possible that future replications and refinements of the methodologies supporting these findings will yield more significant findings and larger effect sizes. The study that provided the strongest evidence for improvement through expert review (Shnabel & Nadler, 2008) provides a suggestive direction for such refinements. The primary revisions to that study involved extensive tailoring and piloting of study materials for new populations. However, this was not the only study whose revisions included this process, reinforcing the possibility that the apparent benefits for this finding occurred by chance. Across all studies, the original findings were statistically

anomalous compared with all replication findings, and the prediction markets, reviewers, and replication teams could not predict which findings would persist with some supporting evidence.

For those findings that failed to improve in replicability, the present understanding of the conditions needed for replicating the effect is not sufficient. This minimally suggests that theoretical revisions are needed in understanding the boundary conditions for observing the effect, and maximally suggests that the original result was a false positive. In the latter case, it is possible that no amount of expertise could have produced a replicable finding. We cannot definitively parse between these possibilities, but the fact that even protocols revised with formal peer review from experts failed to replicate the original effects suggests that theoretical understanding of the findings is too weak to specify replicable conditions (Nosek & Errington, 2020).

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

Conclusion: Is Expertise Irrelevant?

Concluding that expertise is irrelevant for achieving replicable results may be tempting given the very small effect of expert peer review we observed on replication effect sizes. However, that interpretation is unwarranted. The present study is a narrow but important test of the role of expertise in improving replicability. Our control condition was a set of replications using protocols that had mostly failed to replicate in a prior replication project, RP:P. Those original replication protocols were developed in a structured process with original materials and preregistration; replication researchers had sufficient self-identified expertise to design and

conduct the replications; and, designs received informal review by an internal review process and by original authors when they were willing to provide it. This process did not preclude the possibility of errors , but using RP:P protocols meant that the control condition included substantial effort and expertise to conduct a faithful replication. Whether that effort and expertise was sufficient was the open question. The intervention we tested to improve replicability is a function of a particular critique of those failures-to-replicate -- that failure to resolve issues identified by original authors signaled critically problematic features of the replication designs. So, our finding that formal peer review did not systematically improve replicability may be limited to circumstances in which there are already good efforts to conduct high quality replications, such as the variety of systematic replication efforts populating social-behavioral sciences this decade.

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

There may be conditions under which these studies are more replicable, but peer review did not produce them. Peer reviewers were selected for their perceived expertise in the areas of study we investigated. In many cases, the reviewers authored the original research. It is possible, despite the presumed expertise of the reviewers, that they lacked knowledge of what would make

the studies replicable. Other experts may have advised us differently and produced protocols that improved replicability. The current investigation cannot rule out this possibility.

Finally, it is obvious that expertise matters under a variety of conditions and that lack of expertise can have deleterious impacts in specific cases. For example, conducting an eye-tracking study (e.g., Crosby et al., 2008) minimally requires possessing eye-tracking equipment and having sufficient experience with the equipment to operate it properly. Further, replications can fail for technical reasons; experts may be better positioned to identify those technical errors based on experience with instrumentation and protocols. The meaningful question of the role of expertise for replicability is in the zone that replication researchers appear to possess the basic facility for conducting research of that type, and when those replication researchers perceive that they are conducting an effective replication in good faith. That was the circumstance studied in this investigation, and this investigation is hardly the final word.

Action Editor

Daniel J. Simons

Editor

Daniel J. Simons

Author Contributions

CRE and BAN conceived the project and drafted the report. MBM and CRE designed the analysis plan and analyzed the aggregate data. CRE, CRC, JKH, HIJ, IR, MBM, LBL, HR, MC, EB, DB, KSC, and NRB served as team leaders for the sets of replications. DV, CRE, YC, TP, AD, MJ, AS, and BAN designed and analyzed the surveys and prediction markets to elicit peer beliefs. All authors except BAN collected the data. All authors revised and approved the manuscript with two exceptions; sadly, Sebastiaan Pessers and Boban Petrović passed away before the manuscript was finalized.

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.

Chen, S., Szabelska, A., Chartier, C. R., Kekecs, Z., Lynott, D., Bernabeu, P., … Oberzaucher, E. (2018, November 6). Investigating Object Orientation Effects Across 14 Languages. https://doi.org/10.31234/osf.io/t2pjv

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., ... & Zhou, X. (2018). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*, 1-36.

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

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

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., … & Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology, 67*, 68-82.

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

Forsell, E., Viganola, D., Pfeiffer, T., Alemberg, J., Wilson, D., Chen, Y., Nosek, B.A., Johannesson, M., Dreber, A. (2018). Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology*.

Förster, J., Liberman, N., & Kuschel, S. (2008). The effect of global versus local processing styles on assimilation versus contrast in social judgment. *Journal of Personality and Social Psychology*, *94*(4), 579.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(6), 460.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science*, *351*(6277), 1037-1037.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*(1), 1-20.

Harrell Jr, M. F. E. (2019). Package 'Hmisc'. http://cran.r-project.org/web/packages/Hmisc/Hmisc.pdf

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524-532.

LoBue, V., & DeLoache, J. S. (2008). Detecting the snake in the grass: Attention to fear-relevant stimuli by adults and young children. *Psychological Science*, *19*(3), 284-289.

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C. A., Nosek, B. A., Chartier, C. R., … Ratliff, K. A. (2019, December 11). Many Labs 4: Failure to Replicate Mortality Salience Effect With and Without Original Author Involvement. https://doi.org/10.31234/osf.io/vef2c

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... & Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology*, *45*, 142-152.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., … & Nosek, B. A. (2018). Many labs 2: Investigating variation in replicability across sample and setting. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443-490.

Luttrell, A., Petty, R. E., & Xu, M. (2017). Replicating and fixing failed replications: The case of need for cognition and argument quality. *Journal of Experimental Social Psychology, 69,* 178-183.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur?. *Perspectives on Psychological Science, 7*(6), 537-542.

Mathur, M. B. & VanderWeele, T. J. (2020). New statistical metrics for multisite replication projects. *Journal of the Royal Statistical Society: Series A.* https://doi.org/10.1111/rssa.12572

McGuire, W. J. (2004). A perspectivist approach to theory construction. *Personality and Social Psychology Review, 8,* 173-182.

Murray, S. L., Derrick, J. L., Leder, S., & Holmes, J. G. (2008). Balancing connectedness and self-protection goals in close relationships: A levels-of-processing perspective on risk regulation. *Journal of Personality and Social Psychology, 94*(3), 429-459.

Noah, T., Schul, Y., & Mayo, R. (2018). When both the original study and its failed replication are correct: Feeling observed eliminates the facial-feedback effect. *Journal of Personality and Social Psychology, 114*(5), 657-664.

Nosek, B. A., & Errington, T. M. (2017). Reproducibility in cancer biology: making sense of replications. *Elife, 6,* e23383.

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

Nosek, B. A., & Gilbert, E. A. (2016). Let's not mischaracterize the replication studies. *Retraction Watch, 9.*

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

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

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

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

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

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

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

Schwarz, N., & Strack, F. (2014). Does merely going through the same moves make for a "direct" replication? Concepts, contexts, and operationalizations. *Social Psychology, 45*(4), 305-306.

Shnabel, N., & Nadler, A. (2008). A needs-based model of reconciliation: satisfying the differential emotional needs of victim and perpetrator as a key to promoting reconciliation. *Journal of Personality and Social Psychology, 94*(1), 116-132.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359-1366.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science, 12*(6), 1123-1128.

Skorb, L., Aczel, B., Bakos, B., Christ, O., Fedor, A., Feinberg, L., Halasa, E., Jiménez-Leal, W., Kauff, M., Kovacs, M., Krasuska, K. K., Kuchno, K., Manfredi, D., Muda, R., Nave, G., Pękala, E., Pieńkosz, D., Ravid, J., Rentzsch , Katrin, Salamon, J., Schultze, T., Sioma, B., & Hartshorne, J. K. (provisionally accepted). Many Labs 5: Replication Report for Van Dijk, Van Kleef, Steinel, & Van Beest (2008). A social functional approach to emotions in bargaining: When communicating anger pays and when it backfires. Journal of Personality and Social Psychology, 94(4), 600-614. *Advances in Methods and Practices in Psychological Science*.

Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science, 12*(2), 153-156.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, *54*(285), 30-34.

Strack, F. (2016). Reflection on the smiling registered replication report. *Perspectives on Psychological Science, 11*(6), 929-930.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*(1), 59-71.

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

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

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*. https://doi.org/10.1017/S0140525X170019

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).

van Dijk et al., 2008 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).	Payne et al., 2008 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.
Shnabel & Nadler, 2008 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	Albarracin et al. 2008 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

occurred by chance. Nevertheless, if the difference is replicable, then these protocols might help study the role of manipulation checks and effective implementation of the experimental intervention. In this case, the manipulation checks for both protocols suggested that the intervention was effective (Baranski et al., this issue) and yet the outcomes on the dependent variable landed on opposing sides of the statistical significance criterion (*p*'s = .004, .350).	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?

Table 1 - Summary of main protocol differences

Study	Preregistration	Main Differences between RP:P and Revised Protocols
Albarracín et al., Study 5	osf.io/a3pwa/	RP:P protocol collected participants online from MTurk; Revised protocol collected undergraduates in lab.
Albarracín et al., Study 7	osf.io/725ek/	Original authors expressed concern about replicating the study among participants in German because the original materials were validated in English. Both protocols used only English language speaking participants. Additionally, Revised protocol used scrambled sentences to prime instead of word fragments, 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 unfamiliar with experimental scenarios (concerning affirmative action). Revised protocol presented participants with experimental scenarios after they watched a video about affirmative action. RP:P protocol did not include the video about affirmative action.
Forster et al.	osf.io/ev4nv/	The RP:P replication failed at achieving target ambiguity and applicability of stimuli. For the Revised protocol, stimuli were piloted for both aspects at all collection sites; the RP:P protocol used the same stimuli as the previous RP:P replication.
LoBue & DeLoache	osf.io/68za8/	Original authors expressed concerns regarding the physical features of the control stimuli used in the RP:P replication, the age of children recruited, and technical issues such as screen size and software dependent on Internet speed. The Revised protocol used frogs as control stimuli; the RP:P protocol used caterpillars as control stimuli. In addition, the Revised protocol sampled only 3-year-olds along with their parents, (instead of 3-5-year-olds, as in the RP:P protocol). Finally, the study was implemented with internet-independent software (allowing the study to be run offline and therefore not be hampered by internet speed), and on a larger 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 the United States in English
Risen & Gilovich	osf.io/xxf2c/	RP:P protocol recruited subjects on Amazon Mechanical Turk (MTurk) instead of undergraduates at elite universities as in original study. Authors of original study were concerned that MTurk subjects may find the experimental scenarios less personally salient than original sample and may complete experiment while distracted, compromising the cognitive load manipulation. Revised protocol used undergraduates at elite universities.

Shnabel & Nadler	osf.io/q85az/	In the RP:P protocol, participants read a vignette describing an employee who took a 2-week leave from work to go on a honeymoon; in the Revised protocol, participants read a vignette describing a recently unemployed college student who, upon returning from a two-week family visit, was told by his/her roommate that he/she found someone who could commit to paying next year's rent and that the protagonist must move out by the end of the lease. This revision was meant to provide a more relatable experience regarding being the victim or perpetrator of a transgression. The revised materials were created through a pilot study using undergraduate students.
van Dijk et al.	osf.io/xy4ga/	Following the original study, the Revised protocol excluded subjects who had taken prior psychology or economics courses or participated in prior psychology studies. Participants were also situated such that they could not see or hear one another during the experiment. These restrictions were not present in the RP:P protocol.
Vohs & Schooler	osf.io/peuch/	The Revised protocol used different free-will-belief inductions (a rewriting task instead of a reading task, with text from both pulled from the same source) and a revised measure of free-will beliefs (same author team, new instrument) than the RP:P protocol.

Table 2 - Summary of sample sizes and power across studies

	Original Study			RP:P Replication			ML5: RP:P Protocol				ML5: Revised Protocol				
Study	*N*	Power to detect ML5 RP:P	Power to detect ML5 Revised	*N*	Power to detect ML5 RP:P	Power to detect ML5 Revised	Number of Sites	*Total N*	Power to detect original ES	Smallest ES with 90% Power	Number of Sites	*Total N*	Power to detect original ES	Smallest ES with 90% Power	Random assignment to Protocol?
Albarracin et al., Study 5	36	0.06	0.08	88	0.07	0.13	1	580	> 0.99	0.13	8	884	> 0.99	0.11	No
Albarracin et al., Study 7	98	0.05	0.00	105	0.05	0.00	7	878	> 0.99	0.12	7	808	> 0.99	0.12	Yes
Crosby et al.	25	0.39	0.34	30	0.46	0.40	3	140	> 0.99	0.11	3	136	> 0.99	0.11	Yes
Forster et al.	82	0.06	0.07	71	0.05	0.06	8	736	> 0.99	0.13	8	720	> 0.99	0.13	Yes
LoBue & DeLoache	48	0.05	0.06	48	0.05	0.06	4	286	> 0.99	0.19	4	259	> 0.99	0.20	No
Payne et al.	70	0.07	0.00	180	0.10	0.00	4	545	> 0.99	0.14	4	558	> 0.99	0.14	No
Risen & Gilovich	122	0.00	0.00	226	0.00	0.00	1	2811	> 0.99	0.06	4	701	> 0.99	0.12	No
Shnabel & Nadler	94	0.06	0.27	141	0.06	0.40	8	1361	> 0.99	0.05	8	1376	> 0.99	0.05	Yes
van Dijk et al.	103	0.09	0.66	83	0.08	0.56	6	436	> 0.99	0.15	4	119	0.99	0.29	No
Vohs & Schooler	30	0.06	0.06	58	0.06	0.07	4	279	> 0.99	0.19	5	342	> 0.99	0.17	Yes

Note: power calculations used α = .05

Table 3 - Summary of effect sizes across studies

	Original Study			RP:P Replication			ML5: RP:P Protocol			ML5: Revised Protocol		
Study	*N*	*r*	*95% CI*	*N*	*r*	*95% CI*	*N*	*r*	*95% CI*	*N*	*r*	*95% CI*
Albarracin et al., Study 5	36	0.38	.05, .64	88	0.03	-.24, .18	580	0.04	-.04, .12	884	0.09	.03, .14
Albarracin et al., Study 7	98	0.21	.01, .39	5	0.16	-.03, .34	878	0.01	-.19, .21	808	-0.07	-.17, .03
Crosby et al.	25	0.25	.02, .46	30	0.18	-.03, .40	140	0.15	-.01, .30	136	0.14	-.08, .34
Forster et al.	82	0.43	.23, .59	71	0.11	-.13, .34	736	0.03	-.02, .09	720	0.05	-.07, .16
LoBue & DeLoache	48	0.48	.22, .68	48	0.18	-.12, .45	286	0.01	-.19, .21	259	0.04	-.02, .10
Payne et al.	70	0.35	.12, .54	18	0.15	-.00, .29	545	0.05	-.13, .22	558	-0.16	-.44, .15
Risen & Gilovich	12 2	0.19	.01, .36	22 6	0.00	-.13, .13	281 1	-0.04	-.08, -.01	701	0.01	-.13, .11
Shnabel & Nadler	94	0.27	.07, .45	14 1	-0.10	-.27, .07	136 1	0.02	-.03, .08	137 6	0.09	.04, .14
van Dijk et al.	10 3	0.38	.20, .54	83	0.04	-.26, .18	436	0.06	-.06, .18	119	0.23	-.01, .44
Vohs & Schooler	30	0.50	.15, .74	58	0.10	-.17, .35	279	0.04	-.14, .22	342	0.05	-.16, .25

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

Study	*Subset*	*k*	*Estimate (r)*	*p-value*	τ	P_{orig}	*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. Stuy5	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
Payne et al.	RP:P	4	0.05 [-0.13, 0.22]	0.46	0.07	0.03	0.94	75 [0, 100]	50 [0, 100]	0
Payne et al.	Revised	4	-0.16 [-0.44, 0.15]	0.20	0.18	0.03	0.72	25 [0, 50]	0	0
Risen & Gilovich	all	5	-0.04 [-0.14, 0.07]	0.20	0	0.02	0.96	0	0	0
Risen & Gilovich	RP:P	1	-0.04 [-0.08, -0.01]	0.02	0	0.02	0.96	0	0	0
Risen & Gilovich	Revised	4	-0.01 [-0.18, 0.16]	0.87	0.01	0.06	0.83	0	0	0
Shnabel & Nadler	all	16	0.05 [0.02, 0.09]	0.009	0	0.04	0.99	100	0	0
Shnabel & Nadler	RP:P	8	0.02 [-0.03, 0.08]	0.38	0	0.02	0.98	100	0	0
Shnabel & Nadler	Revised	8	0.09 [0.03, 0.14]	0.008	0	0.08	0.98	100	0	0
van Dijk et al.	all	10	0.10 [-0.01, 0.20]	0.07	0.01	0.006	1	100	40 [0, 100]	0
van Dijk et al.	RP:P	6	0.06 [-0.06, 0.19]	0.24	0	0.002	1	100	0	0

van Dijk et al.	Revised	4	0.23 [-0.03, 0.45]	0.07	0	0.19	0.97	100	100	100
Vohs & Schooler	all	9	0.04 [-0.06, 0.15]	0.37	0.06	0.01	0.98	78	22 [0, 100]	0
Vohs & Schooler	RP:P	4	0.04 [-0.15, 0.23]	0.55	0	0.01	0.98	100	0	0
Vohs & Schooler	Revised	5	0.05 [-0.16, 0.25]	0.55	0.11	0.03	0.94	80	20 [0, 100]	0

Study: Name of original study. Subset: all replications for the study, replications using the RP:P protocol, or replications using the Revised protocol. k: number of studies in subset. Estimate (r) and p-value: meta-analytic estimate in replications on Pearson's r scale with 95% confidence interval (brackets) and p-value. : meta-analytic heterogeneity estimate of standard deviation of effects in replications. P_{orig}*: probability that the original study's estimate would be as extreme as actually observed if the original study were consistent with the replications. Probability significance agreement: probability that the meta-analytic estimate in the replications would be statistically significant and would agree in direction with that of the original study if the original and the replications were consistent. Percent above 0, 0.1, and 0.2: estimated percentage of population effects stronger than thresholds of r = 0, 0.1, and 0.2 respectively. Brackets denote 95% confidence intervals, which are omitted for the percentage metrics when they could not be estimated.*

Figure 1 - Power to detect effect of protocol

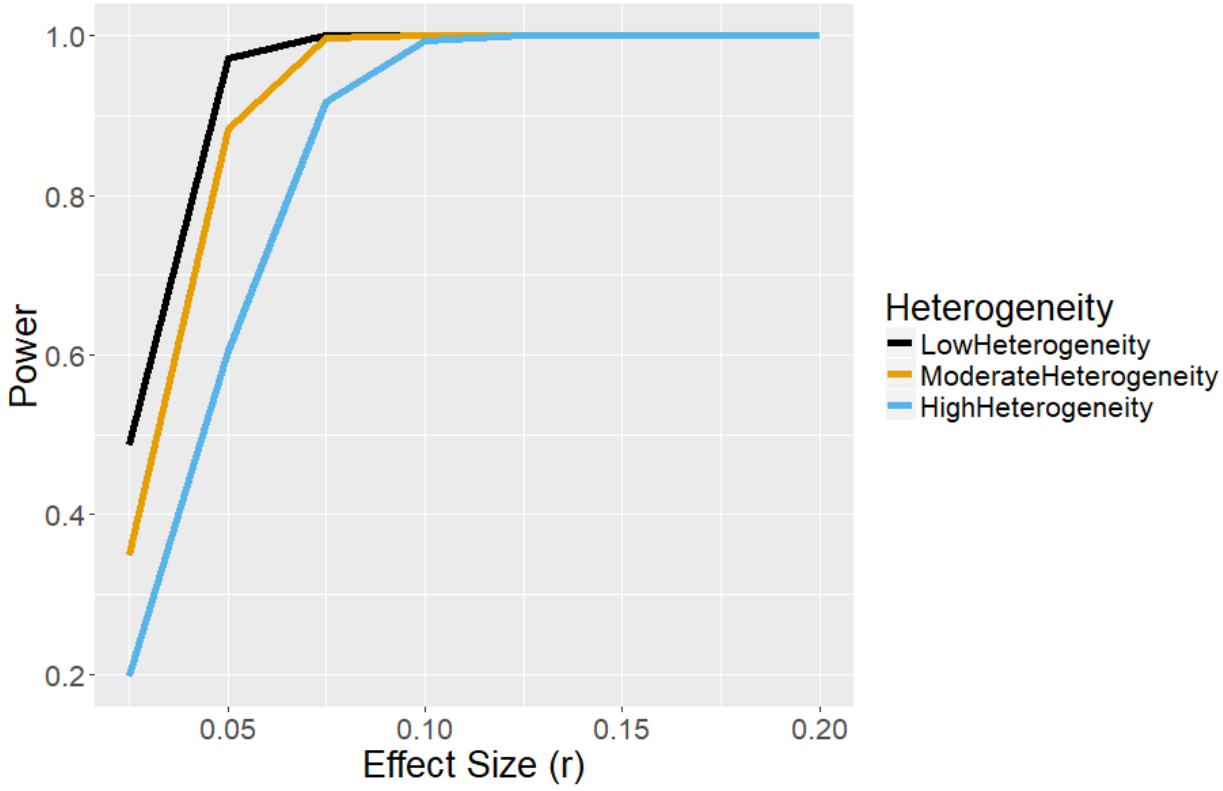
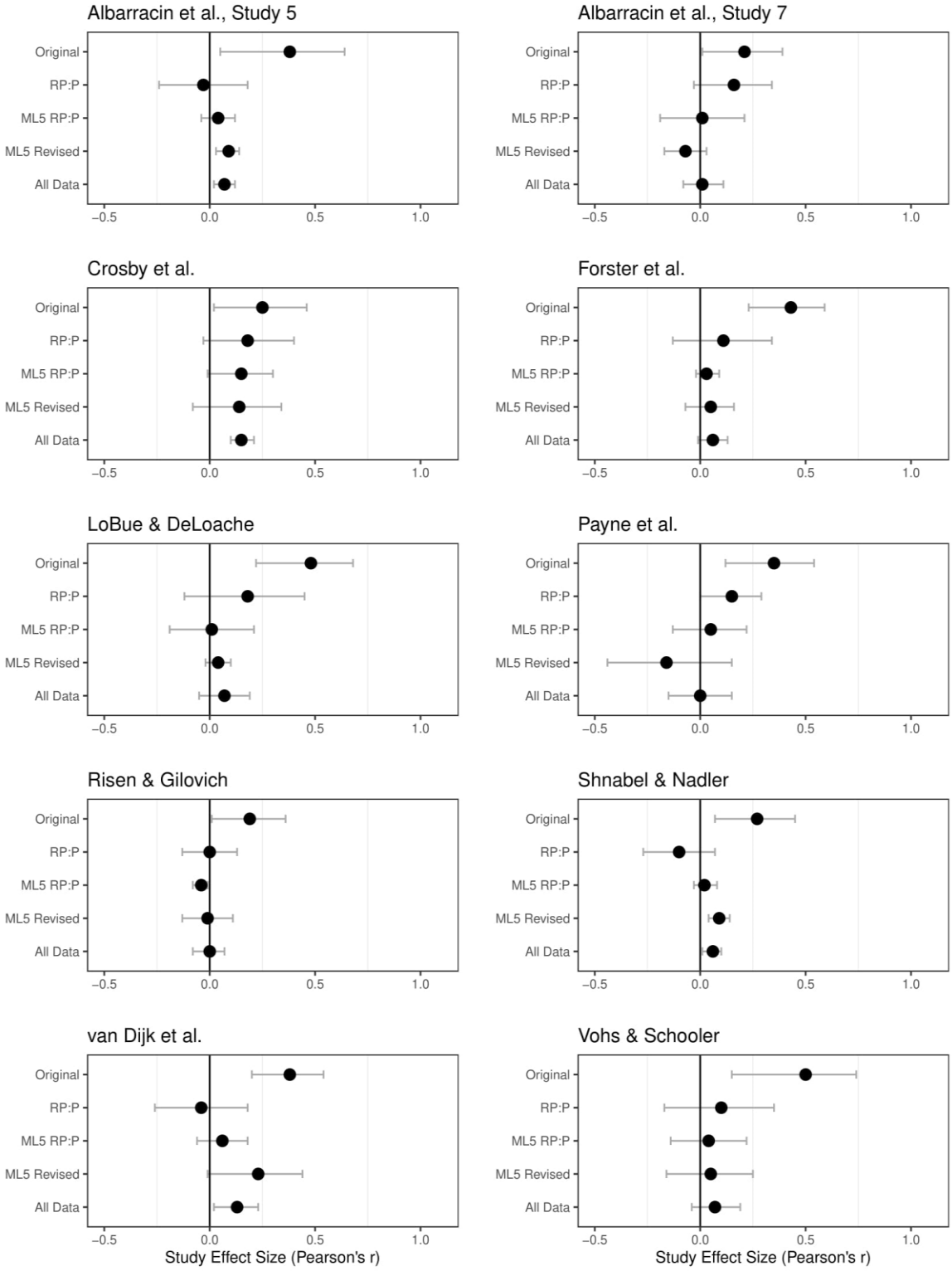
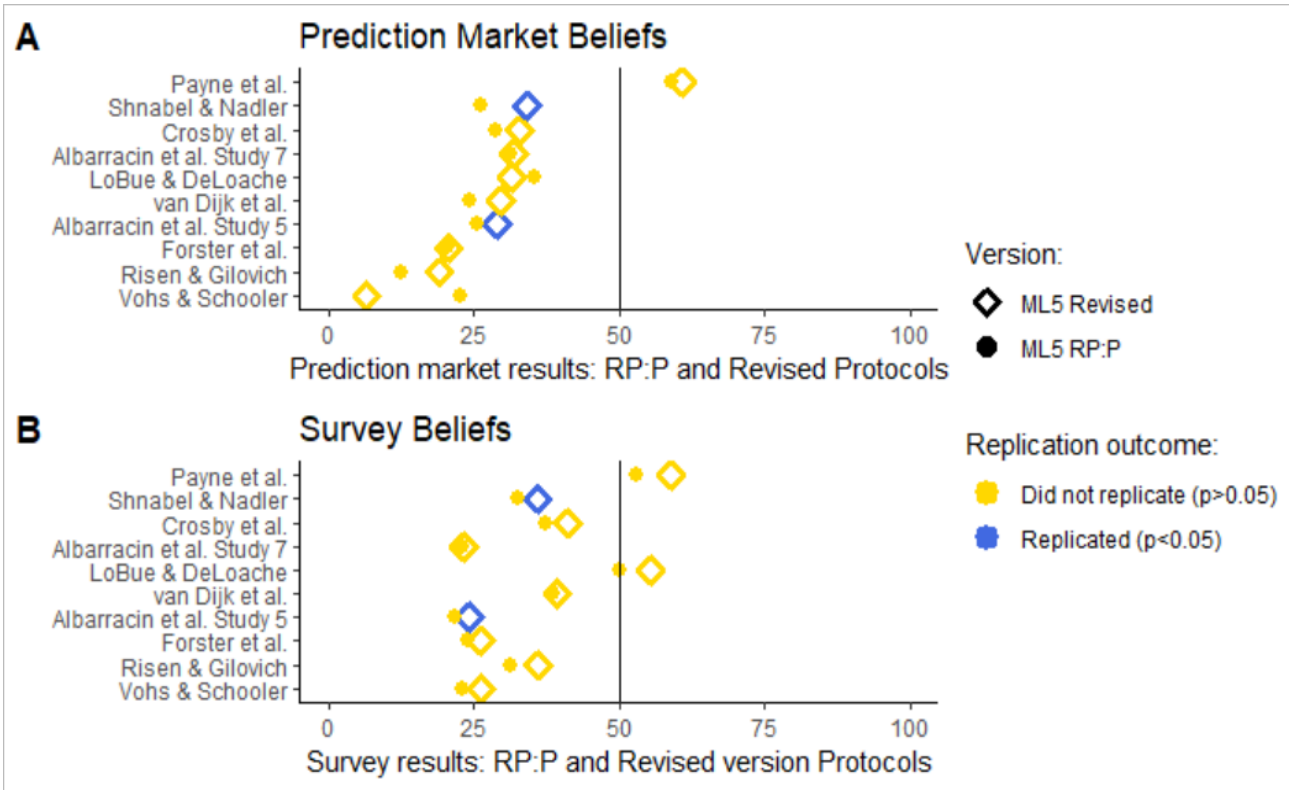


Figure 2 - Effect sizes across study versions



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.

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.

Figure 4 - Effect sizes from individual sites across original studies, RP:P (2015), and Many Labs 5