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, Department of Epidemiology, Harvard University
Erica Baranski, The University of Arizona
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, University of Belgrade
Hugh Rabagliati, Psychology, PPLS, University of Edinburgh
Ivan Ropovik, Faculty of Education, University of Presov
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; California School for Professional Psychology, Alliant International 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, FernUniversität in Hagen, 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, University of Fortaleza, 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
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, 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, University of Belgrade, Serbia
Mark Zrubka, Eötvös Loránd University, Hungary
Brian A. Nosek, Center for Open Science and Department of Psychology, University of Virginia

Abstract

Replication efforts in psychological science sometimes fail to replicate 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 replication 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 on RP:P suggested that lack of adherence to expert review and low-powered tests were the reasons that most of these 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 replication protocols in multiple laboratories (Median number of laboratories per original study = XX; Range XX to YY; Median total sample = XX; Range XX to YY) for high-powered tests of each original finding with both protocols. Overall, XX of 10 RP:P protocols and XX of 10 Revised protocols showed significant evidence in the same direction as the original finding ($p < .05$), compared to an expected XX. The median effect size was [larger/smaller/similar] for Revised protocols (ES = .XX) compared to RP:P protocols (ES = .XX), and [larger/smaller/similar] compared to the original studies (ES = .XX) and [larger/smaller/similar] compared to the original RP:P replications (ES = .XX). Overall, Revised protocols produced [much larger/somewhat larger/similar] effect sizes compared to RP:P protocols (ES = .XX). We also elicited peer beliefs about the replications through prediction markets and surveys of a group of researchers in psychology. The peer researchers predicted that the Revised protocols would

[decrease/not affect/increase] the replication rate, [consistent with/not consistent with] the observed replication results. The results suggest that the lack of replicability of these findings observed in RP:P was [partly/completely/not] due to discrepancies in the RP:P protocols that could be resolved with expert peer review.

Total words = 364

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 criteria, 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 direct replications (Makel et al., 2012).

[1] RP:P included 100 replications, however 3 of the original studies found 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 (Nosek & Errington, 2019; see also 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; Nosek & Errington, 2017; 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; 2019). 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 direct 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 than 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 bulk of our replications under the new, peer-reviewed protocol should now succeed. 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 will be available on the OSF (https://osf.io/7a6rd/). The RP:P Protocols were created from RP:P materials than 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 would like to 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 project. 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, 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.

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 lead(s) 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 RP:P replication teams (https://osf.io/ezcuj/). This protocol was the basis for the RP:P protocol condition. Any differences between the RP:P Protocol and the replication as described in RP:P were minor and reflected practicalities such as lab space, population, climate, and time of year. Next, teams sought out any correspondence and/or responses written by the original authors concerning the RP:P replications. 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 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 the revisions that experts deemed important to make successful replication as likely as possible.

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 differences between the RP:P and Revised protocols. 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 will be 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.[3] 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 = XX.XX, median N = XX.XX, SD = XX.XX; Revised protocols: mean N = XX.XX, median N = XX.XX, SD = XX.XX) 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 $\alpha = .05$ (mean = .XX, range .XX to .XX), 75% of the original effect size with $\alpha = .05$ (mean = .XX, range .XX to .XX), and 50% of the original effect size with $\alpha = .05$ (mean = .XX, range .XX to .XX) for each of the protocols (Camerer et al., 2018). When possible, we randomly assigned participants to one protocol or the other within each data collection site, but randomization was impossible for studies that differed by sample source between protocols (e.g., MTurk vs. in-lab collection). Table 2 provides a summary of

[3] 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).

sample sizes, power, and whether participants were randomly assigned to RP:P or Revised protocols.

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 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

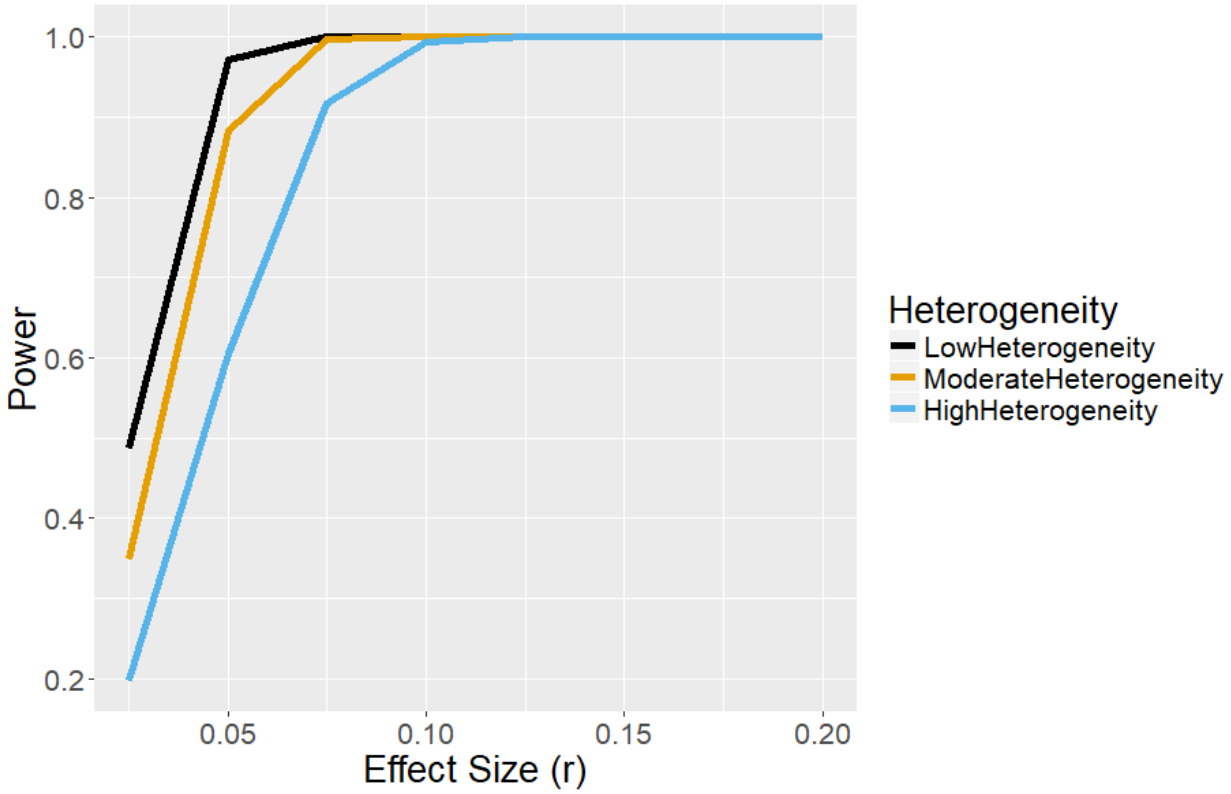
starting from June 21st, 2017, and a total of 31 participants made at least one trade. See the Supplemental Material for more details about the prediction markets and survey.

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.[4] In cases of both low (I^2 = 25%) and moderate (I^2 = 50%) heterogeneity, our minimum planned samples should provide adequate power (> 80%) to detect an 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.

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

Figure 1 - Power to detect effect of protocol



We also simulated our estimated power for our 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.[5] 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 detected relationships between peer beliefs and replication outcomes. The twenty prediction markets would provide 41% power to detect a

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

correlation of .4, 62% power to detect a correlation of .5, 82% power to detect a correlation of .6, and 95% power to detect a correlation of .7. The previous prediction markets have found an average correlation of .58 between peer beliefs and replication outcomes (78% power with twenty markets).

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, XX of 10 replications using the RP:P protocol yielded "statistically significant" aggregated effect sizes ($p < .05$)[6], the median effect size was $r = .XX$, and the median sample size was $N = XX$. Also in the present study, XX of 10 replications using the Revised protocol were statistically significant ($p < .05$), the median effect size was $r = .XX$, and the median sample size was $N = XX$. A full summary of aggregated effect sizes and confidence intervals for each data collection appears in Table 3.

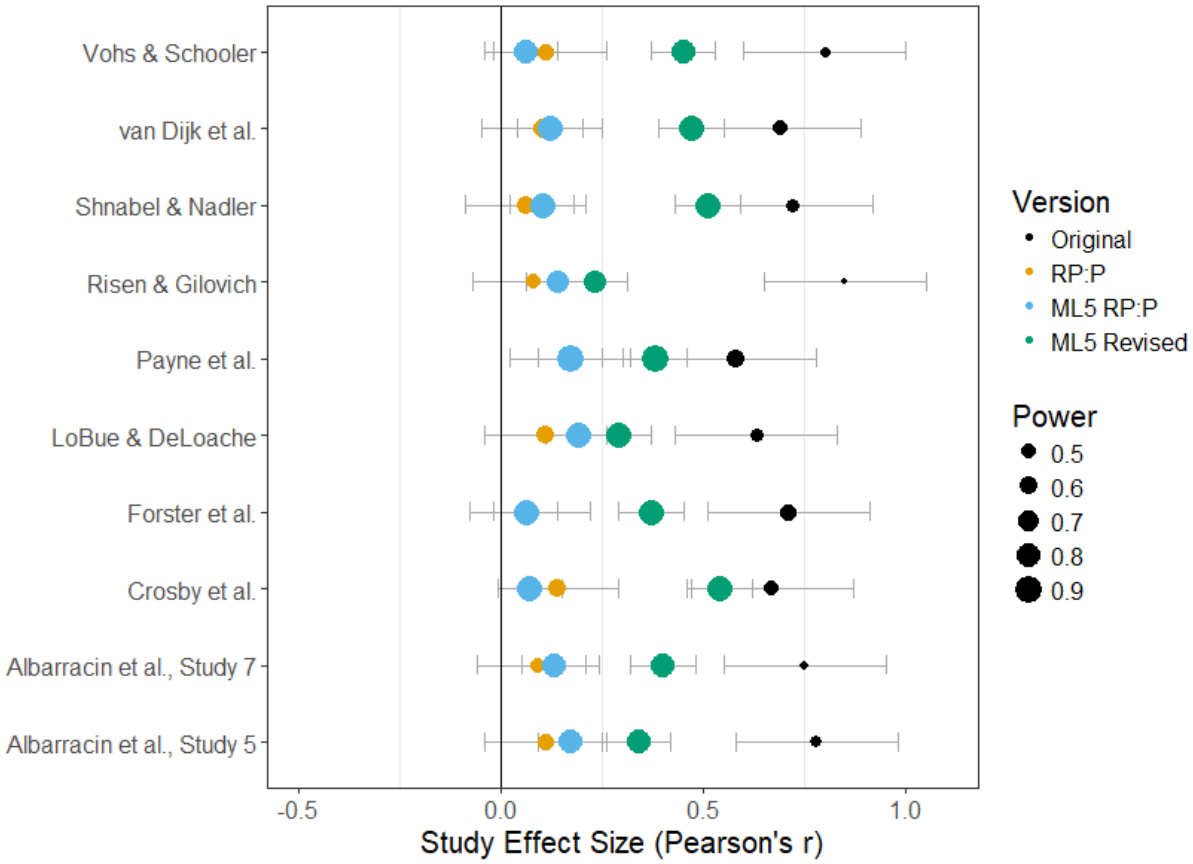
The purpose of this investigation was to test whether a protocol resulting from formal peer review would produce stronger evidence for replicability than a protocol that had not received formal peer review. We tested this in two ways. First, we calculated an effect size for each protocol within each data collection site. Each site implementing both the RP:P protocol

[6] *P*-values and effect sizes taken from the primary tests within each protocol from each individual report.

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 = XX effect sizes[7], with a random intercept of data collection Site (varying from XX to YY depending on Study) nested within Study (10 studies). This model [did/did not] converge so we [did not alter the model/coded Site within Study so that levels of Site were not nested within multiple levels of Study/removed the random intercept of Site]. Then, we added protocol version (RP:P vs. Revised), the hypothesized moderator, as a fixed effect. We found that it had a [large/moderate/small/near zero] effect, b = .XX, SE = .XX, z = .XXX, p = .XXX, 95% CI [.XX, .XX]. That is, effect sizes from Revised protocols were, on average, [larger than/smaller than/similar to] effect sizes from RP:P protocols by b = XX units on the Pearson's r scale. Overall, the effect sizes had [little/a moderate amount/a large amount] of variance not accounted for by the moderator as indexed by Tau = .XX (95% CI [.XX, .XX]) on the Fisher's z scale. Also, the statistic Q suggested [little/moderate/substantial] heterogeneity, Q = XX.XX, p = .XXX, I^2 = XX.XX%.

[7] 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.

Figure 2 - Effect sizes across study versions SIMULATED DATA



Note: For consistency across studies, power is calculated as the likelihood of detecting r = .10. The original studies had an average effect size of r = .37.

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 suggested that these effect sizes [were/were not] reliably different from zero, b = .XX, SE, z = .XXX, p = .XXX, 95% CI [.XX, .XX]. That is, across studies, the Revised protocol point estimates were on average b = .XX units [larger/smaller] than the RP:P point estimate on the Pearson's r scale. Overall, the effect of protocol within each study had [near zero/a little/a moderate amount of/a large amount of] heterogeneity as indicated by Tau = .XX (95% CI [.XX,

.XX]) on the Fisher's *z* scale. And, the Q statistic suggested [little/moderate/substantial] heterogeneity, Q = XX.XX, p = .XXX, I^2 = XX.XX%. XX of the individual studies showed at least a small amount heterogeneity as estimated by Tau being greater than 0.10: XXX (Tau = .XX), XXX (Tau = .XX), and XXX (Tau = .XX).

Exploratory Analyses - Other Evaluations of Replicability

We also examined the cumulative evidence for each of the 10 findings based on the aggregated evidence from the RP:P and Revised protocols. [Note: If there is very little variation between RP:P and Revised versions, then we will examine the cumulative evidence across all forms. If there IS variation between RP:P and Revised versions, then we will look at the cumulative evidence separately for RP:P and Revised versions on each of the outcome criteria. We plan to further examine the evidence for each finding in the following ways. First, we will conduct a form of equivalence testing by assessing, for each study, the proportion of effect sizes in the potentially heterogeneous distribution underlying the replications that are greater than an effect size of 0 and greater than a small effect size (e.g., r = .10; Mathur & VanderWeele, 2017). We will also examine the observed replicability rate (defined as the proportion of replications for each study that estimated a statistically significant point estimate in the same direction as that of the original study) compared to the rate we would expect in ideal circumstances (e.g., those without selective analyses or reporting) following the guidelines of Mathur and VanderWeele (2017). Finally, we will examine the consistency of each set of replications (either separated by protocol or pooled) with the results of the original studies and RP:P replications, again following the guidelines of Mathur and VanderWeele (2017).]

Peer Beliefs

We tested to what extent prediction markets and surveys could successfully predict the replication outcomes. 35 participants 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 on the prediction markets.[8]

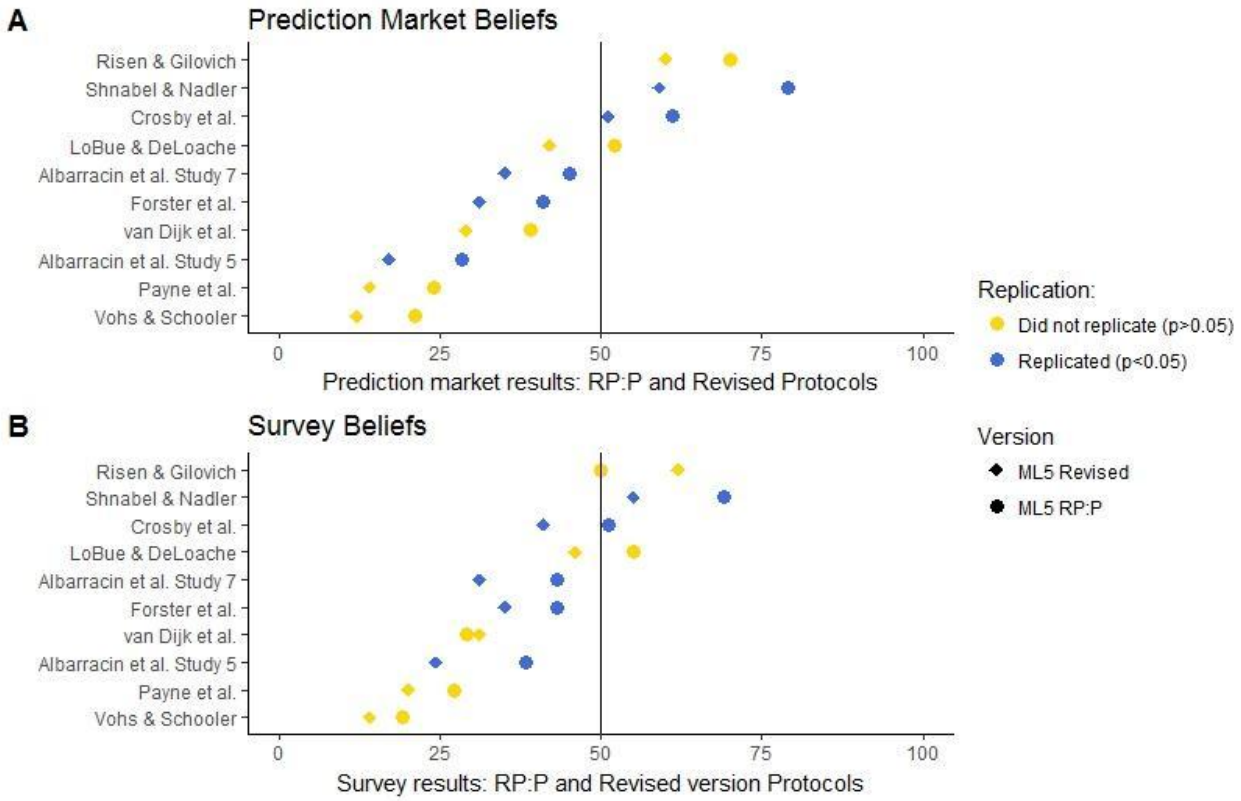
The survey and prediction markets produce a collective peer estimate of the replication success probability for each replication. The mean predicted probability of a statistically significant replication was XX (range XX-XX) for the 10 RP:P protocols and XX (range XX-XX) for the 10 Revised protocols (Wilcoxon signed-ranks test, p = XX, n = 10). The mean survey belief about replication success was XX (range XX-XX) for the 10 RP:P protocols and XX (range XX-XX) for the 10 Revised protocols (Wilcoxon signed-ranks test, p = XX, n = 10). These beliefs about replication success can be compared to the observed replication rate of XX for the 10 RP:P protocols and XX for the 10 Revised protocols.

The relationship between peer beliefs about replication success and replication outcomes are shown in Figure XX, for prediction market beliefs (Panel A) and survey beliefs (Panel B). Both the prediction market beliefs (point-biserial correlation = .XX, p = .XX, n = 20) and the survey beliefs (point-biserial correlation = .XX, p = .XX, n = 20) were [not/weakly/moderately/strongly] correlated with replication outcomes. The prediction market beliefs and survey beliefs were [not/weakly/moderately/strongly] correlated with each other (point-biserial correlation = .XX, p = .XX, 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

[8] The participants that were involved in the replication process of any of the studies analyzed in the Many Lab 5 were not allowed to make predictions on those specific studies, and their answers in the survey about those studies were not used in the following analysis. This exclusion allows us to preserve the consistency in terms of participation between the prediction markets and the survey.

study. If we pool beliefs for the RP:P protocol and the Revised protocol of each study so that we have n = 10 observations (taking the averages of the prediction market prices and survey beliefs for each study), the point-biserial correlation is .XX (p = .XX, n = 10) between the prediction market beliefs and replication outcomes, .XX (p = .XX, n = 10) between the survey beliefs and the replication outcome, and .XX (p = .XX, n = XX) between the prediction market beliefs and the survey beliefs.

Figure 3 - Peer beliefs about replication outcomes SIMULATED DATA



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

Additional Exploratory Analyses

[To be written]

Discussion

[Initial draft, we expect to expand on the points summarized below depending on the observed results and any follow-up exploratory analyses that are conducted]

We tested whether revising protocols based on formal peer review 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 XXXX participants from XX data collection sites, we found that this process of designing replications increased replicability [very little/moderately/substantially] (r = .XX, 95% CI [.XX, .XX], p = .XXX). The results from our Revised protocols were [stronger than/weaker than/similar to] the results from the unaltered RP:P protocols. We also found that researchers in psychology predicted that the Revised protocols would [decrease/not affect/increase] the replication rate, which was [in line with/contrary to] the observed replication results.

[When discussing studies that replicated under the Revised protocol but not RP:P protocol]

For those that failed to replicate with the RP:P protocol, but did replicate with the Revised protocol, we have evidence that incorporating additional expertise in protocol design has a positive impact on improving replicability. This evidence affirms and supports the present theoretical understanding of the conditions necessary to produce the effect, and provides further evidence that some alterations to those conditions negatively impact replicability.

[When discussing studies that did not replicate under either protocol]

For those that failed to improve in replicability with the Revised protocol, we have evidence that the present understanding of the conditions needed for replicating the effect is not sufficient. That 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 an effect. We cannot definitively parse between these possibilities with the present findings, but the fact that even protocols revised with formal peer review from experts failed to replicate the original effects suggests that the phenomena under study need further investigation to identify any conditions under which the hypothesized findings hold and can be replicated.

Specific Implications: The Role of Expertise (and Power) in Replicating 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 [supports, partially supports, does not support] Gilbert et al.'s speculation. There is [strong, some, little, no] evidence that expert feedback contributed to the failures to replicate in RP:P. Also, there is [strong, some, little, no] evidence that the higher powered tests here compared to RP:P increased replication success. Ultimately, whereas RP:P's single replication provided supporting evidence for 1 of 10 findings and conflicting evidence for 9 of 10 findings, the present high-powered, peer reviewed, multiple replications provided supporting evidence for X of 10 findings and conflicting evidence for X of 10 findings. We conclude that expert feedback and lower powered tests are [completely, partly, not at all] sufficient to explain RP:P's failure to replicate 9 of these 10 findings.

[If most/all findings are more replicable with the Revised protocols]

General Implications: The Role of Expertise in Replication

In RP:P, original authors provided feedback *a priori* suggesting potential weaknesses in the replication attempts that could interfere with replicability and reduce observed effect sizes. The present study provides empirical support that their insights were important factors for

achieving replicable findings. This finding suggests that original authors and other domain experts have knowledge about key features of experimental protocols that may not always be evident to other researchers -- even other experts in the same domain.

This finding highlights the importance of communication in the scientific process. Replicability in science is achieved when independent researchers can conduct similar investigations and obtain similar results. If specialized insight is required to achieve replicability, then effective scientific communication will accelerate achieving replicability and advancing scientific knowledge. Such activities include increasing the comprehensiveness of methods sections to capture the key details of the protocol, the addition of constraints on generality sections to papers to know under what conditions the phenomenon is not expected to hold or is unknown (Brandt et al., 2014; Simons, Shoda, & Lindsay, 2017), sharing of materials, data, and code to make recreating the original protocols easier, and encouragement of peer review of methods in advance of data collection, as was done with these studies, with Registered Reports (Chambers, 2013; Nosek & Lakens, 2014).

[If most/all findings are NOT more replicable in the Revised replications]

General Implications: Is Expertise Irrelevant?

Concluding that expertise is irrelevant for achieving replicable results may be tempting given that replicability appears unaffected by expert peer review of replication protocols. However, that interpretation of the results 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 protocols were developed in a structured process with original materials, researchers that had sufficient self-identified expertise to design and conduct the replications,

and with informal review by original authors. As a consequence, the baseline condition already included a great deal of effort and expertise to conduct a faithful replication (whether that effort and expertise was sufficient is an open question). The intervention to improve replicability is a function of a particular critique of those failures-to-replicate -- i.e., that failure to resolve issues identified by original authors signaled critically problematic features of the replication designs. So, the relevance of the lack of impact of formal peer review on improving replicability is limited to this set of circumstances. We found little evidence that conducting formal peer review for these replications can account for their failure to replicate in earlier (or current) replication efforts.

Concluding that conducting formal peer review in advance of conducting studies is not useful for improving methodology and the quality and credibility of findings may also be tempting. That interpretation is also erroneous. A possible reason that we failed to replicate 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. An effect must be replicable under some conditions to test whether different interventions are influential on its replicability. We did not observe any conditions under which these studies were replicable.

Alternatively, there may be conditions under which these studies were 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 may be possible that, despite the presumed expertise of the reviewers, they lacked knowledge of what would make the studies replicable. Other experts may have advised us differently and

produced protocols that did replicate the original study. The current investigation cannot rule out this possibility.

Constraints on Generality [we anticipate that this section will be relevant regardless of results]

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. Overall, the studies were not selected to be representative of any specific field of research. It is therefore unclear as to whether the observed effect of added expertise (or lack thereof) will generalize to other studies.

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 = XX) and studies investigated (10) might not have afforded an adequately powered test of the 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.

[**If some findings became more replicable in the Revised protocols and some did not, then we will include versions of both these sections with appropriate nuance to illustrate how**

both situations can be true. It is likely in the final report that we will still make a nod to both of these even if all findings are or are not improved with the Revised protocols.]

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.

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.

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.

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.

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., 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. (2017). New statistical metrics for multisite replication projects. https://doi.org/10.31219/osf.io/w89s5.

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.

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.

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

Nosek, B. A., & Errington, T. M. (2019). What is replication? https://osf.io/preprints/metaarxiv/u4g6t.

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.

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.

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.

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

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.

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.

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*

Table 1 - Summary of main protocol differences

Study	Preregistration	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*	*N*	Number of Sites	*Total N*	Power to detect original ES	Power to detect 75% of original ES	Power to detect 50% of original ES	Power to detect *r* = .10	Number of Sites	*Total N*	Power to detect original ES	Power to detect 75% of original ES	Power to detect 50% of original ES	Power to detect *r* = .10	Random assignment to Protocol?
Albarracín et al., Study 5	XX	XX		XX						XX					No
Albarracín et al., Study 7	XX	XX		XX						XX					Yes
Crosby et al.	XX	XX		XX						XX					Yes
Forster et al.	XX	XX		XX						XX					Yes
LoBue & DeLoache	XX	XX		XX						XX					No
Payne et al.	XX	XX		XX						XX					No
Risen & Gilovich	XX	XX		XX						XX					No
Shnabel & Nadler	XX	XX		XX						XX					Yes
van Dijk et al.	XX	XX		XX						XX					No
Vohs & Schooler	XX	XX		XX						XX					Yes

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*
Albarracín et al., Study 5	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX
Albarracín et al., Study 7	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX
Crosby et al.	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX
Forster et al.	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX
LoBue & DeLoache	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX
Payne et al.	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX
Risen & Gilovich	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX
Shnabel & Nadler	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX
van Dijk et al.	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX
Vohs & Schooler	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX	XX	.XX	.XX, .XX