Check for updates

Special Section: Many Labs 5 Project
Empirical Article: Registered Report

Advances in Methods and Practices in Psychological Science
2020, Vol. 3(3) 309–331
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2515245920958687
www.psychologicalscience.org/AMPPS

Many Labs 5: Testing Pre-Data-Collection Peer Review as an Intervention to Increase Replicability

Charles R. Ebersole[1], Maya B. Mathur[2], Erica Baranski[3], Diane-Jo Bart-Plange[1], Nicholas R. Buttrick[1], Christopher R. Chartier[4], Katherine S. Corker[5], Martin Corley[6], Joshua K. Hartshorne[7], Hans IJzerman[8,9], Ljiljana B. Lazarević[10,11], Hugh Rabagliati[6], Ivan Ropovik[12,13], Balazs Aczel[14], Lena F. Aeschbach[15], Luca Andrighetto[16], Jack D. Arnal[17], Holly Arrow[18], Peter Babincak[19], Bence E. Bakos[14], Gabriel Baník[19], Ernest Baskin[20], Radomir Belopavlović[21], Michael H. Bernstein[22,23], Michał Białek[24], Nicholas G. Bloxsom[4], Bojana Bodroža[21], Diane B. V. Bonfiglio[4], Leanne Boucher[25], Florian Brühlmann[15], Claudia C. Brumbaugh[26], Erica Casini[27], Yiling Chen[28], Carlo Chiorri[16], William J. Chopik[29], Oliver Christ[30], Antonia M. Ciunci[23], Heather M. Claypool[31], Sean Coary[32], Marija V. Čolić[33], W. Matthew Collins[25], Paul G. Curran[5], Chris R. Day[34], Benjamin Dering[35], Anna Dreber[36,37], John E. Edlund[38], Filipe Falcão[39], Anna Fedor[40], Lily Feinberg[7], Ian R. Ferguson[41,42], Máire Ford[43], Michael C. Frank[44], Emily Fryberger[45], Alexander Garinther[18], Katarzyna Gawryluk[46], Kayla Ashbaugh[47], Mauro Giacomantonio[48], Steffen R. Giessner[49], Jon E. Grahe[45], Rosanna E. Guadagno[50], Ewa Hałasa[51], Peter J. B. Hancock[35], Rias A. Hilliard[52], Joachim Hüffmeier[53], Sean Hughes[54], Katarzyna Idzikowska[46], Michael Inzlicht[55], Alan Jern[52], William Jiménez-Leal[56], Magnus Johannesson[36], Jennifer A. Joy-Gaba[41], Mathias Kauff[57], Danielle J. Kellier[58], Grecia Kessinger[59], Mallory C. Kidwell[60], Amanda M. Kimbrough[61], Josiah P. J. King[6], Vanessa S. Kolb[23], Sabina Kołodziej[46], Marton Kovacs[14], Karolina Krasuska[51], Sue Kraus[62], Lacy E. Krueger[63], Katarzyna Kuchno[51], Caio Ambrosio Lage[64], Eleanor V. Langford[1], Carmel A. Levitan[65], Tiago Jessé Souza de Lima[66], Hause Lin[55], Samuel Lins[39], Jia E. Loy[67], Dylan Manfredi[68], Łukasz Markiewicz[46], Madhavi Menon[25], Brett Mercier[69], Mitchell Metzger[4], Venus Meyet[59], Ailsa E. Millen[35], Jeremy K. Miller[70], Andres Montealegre[71], Don A. Moore[72], Rafał Muda[51], Gideon Nave[68], Austin Lee Nichols[73], Sarah A. Novak[74], Christian Nunnally[75], Ana Orlić[33], Anna Palinkas[14], Angelo Panno[76],

Corresponding Author:
Charles R. Ebersole, University of Virginia, Department of Psychology, 485 McCormick Rd., Charlottesville, VA 22904
E-mail: cebersole@virginia.edu

Kimberly P. Parks[1], Ivana Pedović[77], Emilian Pękala[51], Matthew R. Penner[78], Sebastiaan Pessers[79], Boban Petrović[11,80], Thomas Pfeiffer[81], Damian Pieńkosz[51], Emanuele Preti[27], Danka Purić[11,82], Tiago Ramos[39], Jonathan Ravid[7], Timothy S. Razza[25], Katrin Rentzsch[83], Juliette Richetin[27], Sean C. Rife[84], Anna Dalla Rosa[85], Kaylis Hase Rudy[59], Janos Salamon[14,86], Blair Saunders[87], Przemysław Sawicki[46], Kathleen Schmidt[88], Kurt Schuepfer[31], Thomas Schultze[89,90], Stefan Schulz-Hardt[89,90], Astrid Schütz[91], Ani N. Shabazian[92], Rachel L. Shubella[52], Adam Siegel[93], Rúben Silva[39], Barbara Sioma[51], Lauren Skorb[7], Luana Elayne Cunha de Souza[94], Sara Steegen[79], L. A. R. Stein[22,23,95], R. Weylin Sternglanz[25], Darko Stojilović[96], Daniel Storage[97], Gavin Brent Sullivan[34], Barnabas Szaszi[14], Peter Szecsi[14], Orsolya Szöke[14], Attila Szuts[14], Manuela Thomae[98,99], Natasha D. Tidwell[62], Carly Tocco[26], Ann-Kathrin Torka[53], Francis Tuerlinckx[79], Wolf Vanpaemel[79], Leigh Ann Vaughn[100], Michelangelo Vianello[85], Domenico Viganola[36], Maria Vlachou[79], Ryan J. Walker[31], Sophia C. Weissgerber[101], Aaron L. Wichman[78], Bradford J. Wiggins[59], Daniel Wolf[91], Michael J. Wood[102], David Zealley[59], Iris Žeželj[11,82], Mark Zrubka[103], and Brian A. Nosek[1,104]

[1]Department of Psychology, University of Virginia; [2]Quantitative Sciences Unit, Stanford University; [3]Department of Psychology, The University of Houston; [4]Department of Psychology, Ashland University; [5]Department of Psychology, Grand Valley State University; [6]Psychology, School of Philosophy, Psychology & Language Sciences, University of Edinburgh; [7]Psychology and Neuroscience Department, Boston College; [8]LIP/PC2S, Université Grenoble Alpes; [9]Institut Universitaire de France; [10]Institute of Psychology, Faculty of Philosophy, University of Belgrade; [11]Laboratory for Research of Individual Differences, Faculty of Philosophy, University of Belgrade; [12]Institute for Research and Development of Education, Faculty of Education, Charles University; [13]Faculty of Education, University of Presov; [14]Institute of Psychology, ELTE Eötvös Loránd University; [15]Department of Psychology, University of Basel; [16]Department of Educational Science, University of Genova; [17]Psychology Department, McDaniel College; [18]Department of Psychology, University of Oregon; [19]Institute of Psychology, Faculty of Arts, University of Presov; [20]Department of Food Marketing, Haub School of Business, Saint Joseph's University; [21]Department of Psychology, Faculty of Philosophy, University of Novi Sad; [22]Department of Behavioral and Social Sciences, School of Public Health, Brown University; [23]Department of Psychology, University of Rhode Island; [24]Institute of Psychology, University of Wroclaw; [25]Department of Psychology and Neuroscience, Nova Southeastern University; [26]Department of Psychology, Queens College, City University of New York; [27]Department of Psychology, University of Milano-Bicocca; [28]John A. Paulson School of Engineering and Applied Sciences, Harvard University; [29]Department of Psychology, Michigan State University; [30]Faculty of Psychology, FernUniversität in Hagen; [31]Department of Psychology, Miami University; [32]Quinlan School of Business, Loyola University Chicago; [33]Faculty of Sport and Physical Education, University of Belgrade; [34]Centre for Trust, Peace and Social Relations, Coventry University; [35]Psychology, University of Stirling; [36]Department of Economics, Stockholm School of Economics; [37]Department of Economics, University of Innsbruck; [38]Department of Psychology, Rochester Institute of Technology; [39]Department of Psychology, University of Porto; [40]Institute of Evolution, Centre for Ecological Research, Budapest, Hungary; [41]Department of Psychology, Virginia Commonwealth University; [42]Department of Psychology, New York University; [43]Department of Psychology, Loyola Marymount University; [44]Department of Psychology, Stanford University; [45]Department of Psychology, Pacific Lutheran University; [46]Department of Economic Psychology, Kozminski University; [47]Department of Biology and Biomedical Engineering, Rose-Hulman Institute of Technology; [48]Department of Social & Developmental Psychology, Sapienza University of Rome; [49]Rotterdam School of Management, Erasmus University; [50]Center for International Security and Cooperation, Stanford University; [51]Faculty of Economics, Maria Curie-Skłodowska University; [52]Department of Humanities, Social Sciences, and the Arts, Rose-Hulman Institute of Technology; [53]Department of Psychology, TU Dortmund University; [54]Department of Experimental Clinical and Health Psychology, Ghent University; [55]Department of Psychology, University of Toronto; [56]Department of Psychology, Universidad de los Andes; [57]Department of Psychology, Medical School Hamburg; [58]Perelman School of Medicine, University of Pennsylvania; [59]Department of Psychology, Brigham Young University–Idaho; [60]Department of Psychology, University of Utah; [61]School of Arts, Technology, Emerging Media, & Communication, University of Texas at Dallas; [62]Psychology, Fort Lewis College; [63]Department of

Psychology & Special Education, Texas A&M University-Commerce; [64]Department of Psychology, Pontifical Catholic University of Rio de Janeiro; [65]Department of Cognitive Science, Occidental College; [66]Department of Social and Work Psychology, University of Brasília; [67]Linguistics & English Language, School of Philosophy, Psychology & Language Sciences, University of Edinburgh; [68]Marketing Department, The Wharton School of the University of Pennsylvania; [69]Department of Psychological Science, University of California, Irvine; [70]Department of Psychology, Willamette University; [71]Department of Psychology, Cornell University; [72]Haas School of Business, University of California at Berkeley; [73]Institute of Social and Economic Research, Duy Tan University; [74]Department of Psychology, Hofstra University; [75]Department of Computer Science and Software Engineering, Rose-Hulman Institute of Technology; [76]Department of Human Science, European University of Rome; [77]Department of Psychology, Faculty of Philosophy, University of Niš; [78]Department of Psychological Sciences, Western Kentucky University; [79]Psychology and Educational Sciences, University of Leuven; [80]Institute of Criminological and Sociological Research, Belgrade, Serbia; [81]New Zealand Institute for Advanced Study, Massey University; [82]Department of Psychology, Faculty of Philosophy, University of Belgrade; [83]Psychologische Hochschule Berlin; [84]Department of Psychology, Murray State University; [85]Department of Philosophy, Sociology, Education and Applied Psychology, University of Padova; [86]Doctoral School of Psychology, ELTE Eötvös Loránd University; [87]School of Social Sciences, University of Dundee; [88]School of Psychological and Behavioral Sciences, Southern Illinois University Carbondale; [89]Institute of Psychology, University of Göttingen; [90]Leibniz Science Campus Primate Cognition, Göttingen, Germany; [91]Department of Psychology, University of Bamberg; [92]School of Education, Loyola Marymount University; [93]Cultivate Labs, Chicago, Illinois; [94]Department of Psychology, University of Fortaleza; [95]Rhode Island Training School, Rhode Island Department of Children, Youth and Families; [96]Belgrade, Serbia; [97]Department of Psychology, University of Denver; [98]MEU - Die Multiversität; [99]Diploma University of Applied Sciences; [100]Department of Psychology, Ithaca College; [101]Department of Psychology, University of Kassel; [102]Department of Psychology, University of Winchester; [103]Department of Psychology, University of Amsterdam; and [104]Center for Open Science, Charlottesville, Virginia

Abstract

Replication studies in psychological science sometimes fail to reproduce prior findings. If these studies 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 replication studies from the Reproducibility Project: Psychology (RP:P; Open Science Collaboration, 2015) for which the original authors had expressed concerns about the replication designs before data collection; only one of these studies had yielded a statistically significant effect ($p < .05$). 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 the original effects. We revised the replication protocols and received formal peer review prior to conducting new replication studies. We administered the RP:P and revised protocols in multiple laboratories (median number of laboratories per original study = 6.5, range = 3–9; median total sample = 1,279.5, range = 276–3,512) for high-powered tests of each original finding with both protocols. Overall, following the preregistered analysis plan, we found that the revised protocols produced effect sizes similar to those of the RP:P protocols ($\Delta r = .002$ or .014, depending on analytic approach). The median effect size for the revised protocols ($r = .05$) was similar to that of the RP:P protocols ($r = .04$) and the original RP:P replications ($r = .11$), and smaller than that of the original studies ($r = .37$). Analysis of the cumulative evidence across the original studies and the corresponding three replication attempts provided very precise estimates of the 10 tested effects and indicated that their effect sizes (median $r = .07$, range = .00–.15) were 78% smaller, on average, than the original effect sizes (median $r = .37$, range = .19–.50).

Keywords

replication, reproducibility, metascience, peer review, Registered Reports, open data, preregistered

Received 12/7/18; Revision accepted 8/21/20

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 be replicated—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, multistudy projects have failed to replicate a substantial portion of the published findings that they tested. For example, success rates (based on each project's primary replication criterion) have been as follows: 10 of 13 findings (77%) in Klein et al. (2014), 36 of 97 findings (37%) in the Reproducibility Project: Psychology (RP:P; Open Science Collaboration, 2015)[1], 11 of 18 findings (61%) in Camerer et al. (2016), 3 of 10 findings (30%) in Ebersole et al. (2016), 29 of 37 findings (78%) in Cova et al. (2018), 13 of 21 findings (62%) in Camerer et al. (2018), and 14 of 28 findings (50%) in Klein et al. (2018). Moreover, replication studies, even when finding supporting evidence for the original claim (e.g., $p < .05$), tend to yield a smaller observed effect size compared with 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 replication studies (Makel, Plucker, & Hegarty, 2012).

A finding might not be replicated for several reasons. The initial finding might have been a false positive, reflecting either a "normal" Type I error or one made more likely by selectively reporting 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, Loewenstein, & Prelec, 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 be 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 original authors to provide feedback on study protocols before they were administered. Nevertheless, these design efforts may not have been sufficient to reduce or eliminate false negatives. For example, in the 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," which means that, a priori, the original authors had identified potential shortcomings 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 with a much smaller effect size than in the original study. In a critique of the RP:P (Gilbert, King, Pettigrew, & Wilson, 2016), these unresolved issues were cited as a likely explanation for replication failure (but see responses by Anderson et al., 2016; Nosek & Gilbert, 2016).

Unfaithful or Ineffective Methods as a Moderator of Replicability

A replication study is an attempt to reproduce a previously observed finding with no a priori expectation for a different outcome (see Nosek & Errington, 2017, 2020; Zwaan, Etz, Lucas, & Donnellan, 2018). Nevertheless, a replication study may still produce a different outcome for a variety of reasons (Gilbert et al., 2016; Luttrell, Petty, & Xu, 2017; Noah, Schul, & Mayo, 2018; Open Science Collaboration, 2015; Petty & Cacioppo, 2016; Schwarz & Strack, 2014; Strack, 2016; Stroebe & Strack, 2014). 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 et al.'s, 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 study are theory laden (Nosek & Errington, 2017, 2020). Because 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 study are theoretically irrelevant for observing the identified effect.

As is true for 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 the RP:P that were identified as not endorsed by original authors offer a unique opportunity to test this hypothesis. These RP:P protocols were deemed by the replication teams to be replications of the original studies, but the original authors expressed concerns prior to data collection. Therefore, if any failed replications can be explained as due to poor replication design, these are among the top candidates. Thus, we revised 10 of the 11 nonendorsed protocols from the RP:P and subjected them to peer review before data collection, using the Registered Report model (Center for Open Science, n.d.; Chambers, 2013; Nosek & Lakens, 2014). Once the protocols were accepted following formal peer review, they were preregistered on the Open Science Framework (OSF; see Table 1). Then, we conducted replications using both the RP:P protocols and the revised protocols; for each original study, multiple laboratories contributed 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 protocols would be expected to improve replicability and increase effect sizes to be closer to those of the original studies. This would not necessarily mean that *all* failures to replicate are due to poor design. After all, our sample of studies was chosen because they are among the most likely published replications to have faulty designs. However, such an outcome would suggest that published replicability rates are overly pessimistic. Note that the replications using the original RP:P protocols served as a control: If we found that both protocols led to successful replications, then the failures in the RP:P were more likely due to low power or some unexpected difference in the replication teams themselves. In contrast, if most of the replication studies failed even after expert input, this would cast 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 a survey and prediction markets to bet on whether the RP:P and revised protocols would successfully replicate the original findings. Prior evidence from surveys and prediction markets suggests that researchers can effectively anticipate replication success or failure (Camerer et al., 2016, 2018; Dreber et al., 2015; Forsell et al., 2019). Thus, this parallel effort provided an opportunity to test whether researchers anticipated improved replicability with the 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

Preregistration

The design and confirmatory analyses were preregistered on OSF (https://osf.io/nkmc4/). Links to the preregistrations for the individual replication studies can be found in Table 1.

Data, materials, and online resources

All data and code are available on OSF (https://osf.io/7a6rd/). The RP:P protocols were created from the original RP:P materials, which can be found at https://osf.io/ezcuj/. The Supplemental Material (http://journals.sagepub.com/doi/suppl/10.1177/2515245920958687) contains methodological information about the additional measures of replicability and about the prediction market, as well as additional results.

Reporting

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

Ethical approval

Data were collected in accordance with the Declaration of Helsinki. Ethics approval for individual studies was given by local institutional review boards for all data-collection sites.

Method

The RP:P studies we selected for replication were those labeled "not endorsed" (Open Science Collaboration, 2015). For each of the 11 candidate studies, we sought one or more team leads to conduct the new replications and enough research teams to satisfy our sampling plan (discussed later in this section). We recruited researchers through professional listservs, personal contacts, and

Table 1. Summary of the Main Protocol Differences

Original study	Preregistration	Main differences between the Reproducibility Project: Psychology (RP:P) protocol and the revised protocol in Many Labs 5 (ML5)
Albarracín et al. (2008), Experiment 5	osf.io/6qn4t/	In the ML5 RP:P protocol, participants were Amazon Mechanical Turk workers who completed the experiment online; in the revised protocol, participants were undergraduates who were tested in the lab.
Albarracín et al. (2008), Experiment 7	osf.io/725ek/	The original authors expressed concern that the RP:P replication study was conducted in German because the original materials were validated in English. Both ML5 protocols used only English-speaking participants. Additionally, the revised protocol used scrambled sentences instead of word fragments to prime goals, because word fragments did not often elicit target words in the RP:P replication study. The ML5 RP:P protocol used word fragments.
Crosby, Monin, & Richardson (2008)	osf.io/tj6qh/	The original authors were concerned that participants in the RP:P replication study might have been unfamiliar with the experimental scenarios (concerning affirmative action). In the ML5 revised protocol, participants were presented with the experimental scenarios after they watched a video about affirmative action. The ML5 RP:P protocol did not include the video about affirmative action.
Förster, Liberman, & Kuschel (2008)	osf.io/ev4nv/	The RP:P replication study failed at achieving target ambiguity and applicability of stimuli. In ML5, stimuli for the revised protocol were pilot-tested for both aspects; the RP:P protocol used the same stimuli as the previous RP:P replication study.
LoBue & DeLoache (2008)	osf.io/68za8/	The original authors expressed concerns regarding the physical features of the control stimuli used in the RP:P replication study, the age of children recruited, and technical issues such as screen size and software dependent on Internet speed. In ML5, 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- to 5-year-olds, as in the RP:P protocol. Finally, the revised protocol was implemented with Internet-independent software (which allowed the study to be run offline and therefore not hampered by Internet speed), and on a larger screen, more similar to those used in the original studies.
Payne, Burkley, & Stokes (2008)	osf.io/4f5zp/	In the ML5 RP:P protocol, data were collected at sites in Italy with materials written in Italian; in the revised protocol, data were collected at sites in the United States with materials written in English.
Risen & Gilovich (2008)	osf.io/xxf2c/	In the RP:P replication study, participants were Amazon Mechanical Turk workers, but in the original study, participants were undergraduates at elite universities. The authors of the original study were concerned that Mechanical Turk workers might find the experimental scenarios less personally salient than the original sample did and might complete the experiment while distracted, compromising the cognitive-load manipulation. The ML5 revised protocol used undergraduates at elite universities; the RP:P protocol used Mechanical Turk workers.
Shnabel & Nadler (2008)	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 2-week family visit, was told by his or her roommate that he or she had to move out by the end of the lease because the roommate had found someone who could commit to paying the next year's rent. 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, van Kleef, Steinel, & van Beest (2008)	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 (2008)	osf.io/peuch/	The revised protocol used a different free-will-belief induction than the RP:P protocol did (a rewriting task instead of a reading task; text in the two protocols was pulled from the same source). Also, the revised protocol used a revised measure of free-will beliefs.

StudySwap (https://osf.io/view/StudySwap/). We were able to satisfy our recruitment goals for 10 of the 11 replication studies (all except Murray, Derrick, Leder, & Holmes, 2008). For each of the 10 studies, we conducted two replications: one using the RP:P protocol and the other using the revised protocol that was approved following formal peer review. Because the 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/). Differences between our RP:P protocol and the replication as described in the RP:P reflected practicalities such as lab space, population, climate, and time of year (see the other Many Labs 5 articles in this issue for details of the RP:P replications). Next, teams sought out any correspondence 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. These revisions were the basis for our revised protocols. 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 Editor's agreement 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. On the basis of the 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 the 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 the original studies in that participants were recruited via Mechanical Turk or from different countries or in that participants of different age ranges were recruited). 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. We were agnostic as to which types of changes would be most likely to yield successful replications. We sought to implement 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 our RP:P and revised protocols (i.e., the primary changes suggested by reviewers and previous correspondence). The reports for 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, we drafted the present summary report and submitted it to peer review before the two project organizers knew the results of the majority of the replications (B. A. Nosek knew none of the results; C. R. Ebersole was directly involved with data collection for two of the sets of replications and was aware of only those results). The two project organizers had primary responsibility for drafting the manuscript, and the other authors contributed to revisions, knowing the outcomes of at most one set of replications during the writing process (depending on which individual studies they helped conduct). The full reports of the individual replication studies are reported separately in this issue (Baranski et al., 2020; Buttrick et al., 2020; Chartier et al., 2020; Corker et al., 2020; Ebersole et al., 2020; IJzerman et al., 2020; Lazarević et al., 2020; Mathur et al., 2020; Rabagliati et al., 2020; Skorb et al., 2020).

Sampling plan

We collected data for 20 protocols in total—2 protocols (RP:P and revised) for each of 10 original studies.[4] For each protocol, we sought a minimum of three data-collection sites unless the study sampled from Mechanical Turk (e.g., the RP:P protocol of Risen & Gilovich, 2008). At each site and for each protocol, 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 for a protocol would be difficult to achieve 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 (median number of laboratories per original study = 6.5, range = 3–9; median total sample = 1,279.5, range = 276–3,512; 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 replication studies (mean N = 103, median N = 85.5, SD = 61.94). 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 (see Table 2); for the other half, randomization was impossible because of the revisions to the RP:P protocol (e.g., data collection on Mechanical Turk vs. in the lab).

Eliciting peer beliefs

Predictions about replication success guided the selection and revision of original studies in this project. To assess whether other researchers shared these predictions, we measured peer beliefs about the replication protocols. Following previous efforts (Camerer et al., 2016, 2018; Dreber et al., 2015; Forsell et al., 2019), we invited psychology researchers to predict the replication outcomes for the 10 RP:P protocols and 10 revised protocols in prediction markets and a survey. Before being allowed to trade in the markets, participants had to complete a survey in which they rated the probability of successful replication (a statistically significant effect, p <.05, in the same direction as in the original study) for each of the 20 protocols. In the prediction market, participants traded contracts worth money if the original study's effect was replicated and worth nothing if it was not replicated. With some caveats (Manski, 2006), the prices of such contracts can be interpreted as the probabilities that the market assigns to successful replication. For each study, participants could enter the quantity of the contract they wanted to buy (if they believed that the true probability that the effect would be replicated was higher than the one specified by the current price) or to sell (if they believed that the true probability that the effect would be replicated was 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 Many Labs 5 RP:P protocol, the revised protocol, and a document summarizing the differences between the two. They were informed that all the replication studies had at least 80% statistical power. The prediction markets were open for 2 weeks starting June 21, 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 original studies' effects when the studies were replicated using protocols from the RP:P and when they were replicated using protocols revised through expert peer review. We calculated our power to detect an effect of protocol within each set of studies (k = 10). The results are displayed in Figure 1 (see https://osf.io/j5vnh/ for the scripts for the power analysis and figure). In cases of both low (I^2 = 25%) and moderate (I^2 = 50%) heterogeneity, our minimum planned samples should have provided 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 have provided adequate power to detect an effect of protocol as small as r = .075. Power under all heterogeneity assumptions approached 100% for effects with an r value of .10 or more. As a comparison, the difference in r values between effect sizes reported in the original studies and those reported in the RP:P was, on average, .27. At relatively high heterogeneity (I^2 = 73–75%), our minimum planned sample would achieve adequate power (> 80%) at an average effect-size difference of .125 between protocols.

We also simulated our estimated power for a second analysis strategy, that being meta-analyzing the effect size from each protocol within each individual site and testing protocol version as a meta-analytic moderator (see https://osf.io/dhr3p/ for the power simulation script). These power estimates were slightly lower. At relatively high heterogeneity (I^2 = 73–75%), our minimum planned sample would achieve adequate power (90%) at an average effect-size difference of .125 between protocols. However, it is worth noting that both sets of power analyses relied on assumptions about the amounts of different sources of heterogeneity (see https://osf.io/dhr3p/ for the power simulation script).

Finally, we estimated power for detecting relationships between peer beliefs and replication outcomes. The 20 prediction markets provided 41% power to detect a 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 (Camerer et al., 2016, 2018; Dreber et al., 2015; Forsell et al., 2019) found an average correlation of .58 between peer beliefs and replication outcomes (78% power with 20 markets).

Results

Confirmatory analyses: comparing results from the RP:P and revised protocols

We replicated each of 10 studies with two large-sample protocols, one based on the RP:P replication study

Table 2. Summary of Sample Sizes and Power in the Many Labs 5 (ML5) Studies

							ML5								
	Original study			RP:P replication			ML5: RP:P protocol				ML5: revised protocol				
Study	*N*	Power to detect ML5 RP:P protocol's ES	Power to detect ML5 revised protocol's ES	*N*	Power to detect ML5 RP:P protocol's ES	Power to detect ML5 revised protocol's ES	Number of sites	Total analysis *N*	Power to detect original ES	Smallest ES for 90% power	Number of sites	Total analysis *N*	Power to detect original ES	Smallest ES for 90% power	Random assignment to protocol?
Albarracín et al. (2008), Experiment 5	36	.06	.08	88	.07	.13	1	580	> .99	0.13	8	884	> .99	0.11	No
Albarracín et al. (2008), Experiment 7	98	.05	.00	105	.05	.00	7	878	> .99	0.12	7	808	> .99	0.12	Yes
Crosby, Monin, & Richardson (2008)	25	.39	.34	30	.46	.40	3	140	> .99	0.11	3	136	> .99	0.11	Yes
Förster, Liberman, & Kuschel (2008)	82	.06	.07	71	.05	.06	8	736	> .99	0.13	8	720	> .99	0.13	Yes
LoBue & DeLoache (2008)	48	.05	.06	48	.05	.06	4	286	> .99	0.19	4	259	> .99	0.20	No
Payne, Burkley, & Stokes (2008)	70	.07	.00	180	.10	.00	4	545	> .99	0.14	4	558	> .99	0.14	No
Risen & Gilovich (2008)	122	.00	.00	226	.00	.00	1	2,811	> .99	0.06	4	701	> .99	0.12	No
Shnabel & Nadler (2008)	94	.06	.27	141	.06	.40	8	1,361	> .99	0.05	8	1,376	> .99	0.05	Yes
van Dijk, van Kleef, Steinel, & van Beest (2008)	103	.09	.66	83	.08	.56	6	436	> .99	0.15	4	119	.99	0.29	No
Vohs & Schooler (2008)	30	.06	.06	58	.06	.07	4	279	> .99	0.19	5	342	> .99	0.17	Yes

Note: The power calculations used $\alpha = .05$. ES = effect size (Pearson's *r*); RP:P = Reproducibility Project: Psychology.

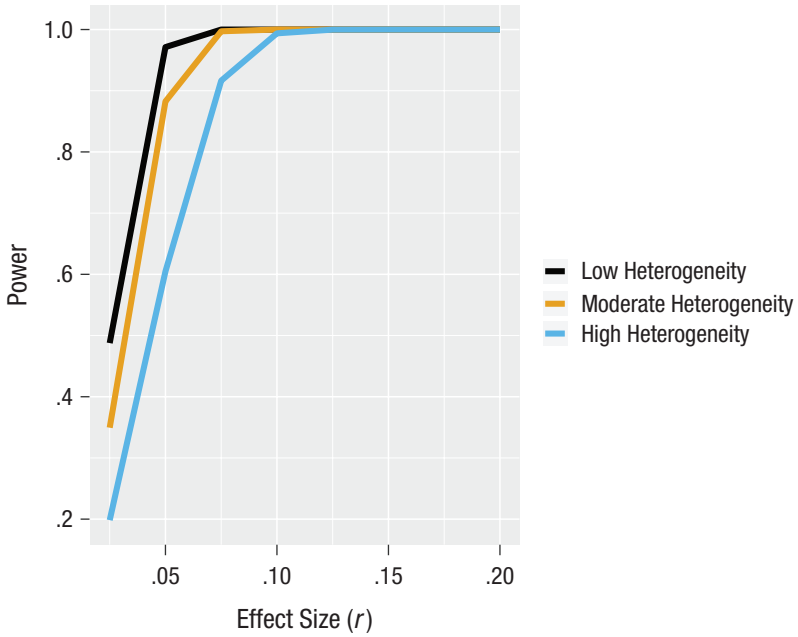

Fig. 1. Power to detect an effect of protocol on the effect sizes obtained within each set of Many Labs 5 replication studies, given low (25%), medium (50%), and high (75%) heterogeneity.

(Open Science Collaboration, 2015) and another that was revised on the basis of formal peer review by experts. In the original reports, all 10 key findings were statistically significant ($p < .05$); the median effect size, measured as r, was .37; and the median sample size was 76. In the RP:P, 1 of the 10 effects was statistically significant ($p < .05$), the median effect size was .11, and the median sample size was 85.5.

In the present study, none of the 10 replications using the RP:P protocol yielded a statistically significant meta-analytic effect size ($p < .05$), the median effect size was .04, and the median sample size was 562.5. Also, 2 of the 10 replications[5] using the revised protocol yielded statistically significant meta-analytic effect sizes ($p < .05$), the median effect size was .07, and the median sample size was 629.5. Gauging replication success on the basis of whether the observed effects are statistically significant is subject to substantial caveats. For example, depending on the power of the original study and the replication studies, the expected proportion of significant effects in the replication studies can be quite low even when the original effect is consistent with the effects observed in the replication studies (Andrews & Kasy, 2019; Patil, Peng, & Leek, 2016). Table 3 presents a full summary of aggregated effect sizes and confidence intervals for the original studies, their corresponding RP:P replication studies, and the two protocols in the current project. As a benchmark to help us 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 replication study was consistent with the original study (Mathur & VanderWeele, 2020).

The purpose of this investigation was to test whether protocols resulting from formal peer review would produce stronger evidence for replicability than protocols 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 101 effect sizes,[6] 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, $b = 0.002$, 95% confidence interval (CI) = [–.04, .04], $SE = 0.02$, $z = 0.091$, $p = .928$. That is, effect sizes from revised protocols were, on average, 0.002 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, $\tau = .05$ (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, $Q = 147.07$, $p = .001$, $I^2 = 26.57\%$.

Table 3. Summary of Effect Sizes Across Studies

	Original study		RP:P replication		ML5: RP:P protocol		ML5: revised protocol	
Original study	*N*	*r*	*N*	*r*	*N*	*r*	*N*	*r*
Albarracín et al. (2008), Experiment 5	36	.38 [.05, .64]	88	−.03 [−.24, .18]	580	.04 [−.04, .12]	884	.09 [.03, .14]
Albarracín et al. (2008), Experiment 7	98	.21 [.01, .39]	105	.16 [−.03, .34]	878	.01 [−.19, .21]	808	−.07 [−.17, .03]
Crosby, Monin, & Richardson (2008)	25	.25 [.02, .46]	30	.18 [−.03, .40]	140	.15 [−.01, .30]	136	.14 [−.08, .34]
Förster, Liberman, & Kuschel (2008)	82	.43 [.23, .59]	71	.11 [−13, .34]	736	.03 [−.02, .09]	720	.05 [−.07, .16]
LoBue & DeLoache (2008)	48	.48 [.22, .70]	48	.18 [−.10, .46]	286	.01 [−.19, .21]	259	.04 [−.02, .10]
Payne, Burkley, & Stokes (2008)	70	.35 [.12, .54]	180	.15 [.00, .29]	545	.05 [−.13, .22]	558	−.16 [−.44, .15]
Risen & Gilovich (2008)	122	.19 [.01, .36]	226	.00 [−.13, .13]	2,811	−.04 [−.08, −.01]	701	−.01 [−.13, .11]
Shnabel & Nadler (2008)	94	.27 [.07, .45]	141	−.10 [−.27, .07]	1,361	.02 [−.03, .08]	1,376	.09 [.04, .14]
van Dijk, van Kleef, Steinel, & van Beest (2008)	103	.38 [.20, .54]	83	−.04 [−.26, .18]	436	.06 [−.06, .18]	119	.23 [−.01, .44]
Vohs & Schooler (2008)	30	.50 [.15, .74]	58	.10 [−.17, .35]	279	.04 [−14, .22]	342	.05 [−.16, .25]

Note: Values in brackets are 95% confidence intervals. RP:P = Reproducibility Project: Psychology; ML5 = Many Labs 5.

For the second test, we conducted a random-effects meta-analysis on the estimates of the effect of protocol within each replication study. 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 10 estimates suggested that these effect sizes were not reliably different from zero, $b = 0.014$, 95% CI = [−.02, .05], $SE = 0.01$, $t = 0.968$, $p = .335$. Across studies, the point estimates for revised protocols were thus, on average, 0.014 units larger than the point estimate for RP:P protocols on the Pearson's *r* scale. Overall, the effect of protocol within each study had a fairly small amount of heterogeneity, $\tau = .034$ (95% CI = [0, .06]) on the Fisher's *z* scale. However, the *Q* statistic suggested significant heterogeneity, $Q = 21.81$, $p = .010$, $I^2 = 60.89\%$. Collapsing the data across protocols, we found that only one of the individual studies (Ebersole et al.'s, 2020, replication of Payne, Burkley, & Stokes, 2008) showed at least a small amount of heterogeneity, as indicated by a τ value greater than .10 ($\tau = .16$ for this study)

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 yielded very weak effect sizes with narrow confidence intervals. Nevertheless, two of the revised protocols showed effects below the $p < .05$ threshold (p values of .009 and .005), whereas none of the RP:P protocols did so. Although this pattern might appear to support the hypothesis that expert peer review could improve replicability, 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 the replicability of findings more than trivially, but perhaps it did for these two studies? Of the two statistically significant effects obtained with the revised protocol, the observed effect sizes were 76% and 67% smaller than the those reported for the original studies. Comparing the RP:P and revised protocols for each of these findings indicated that for only one of the two tests was the revised protocol's effect size significantly larger ($p = .601$ for Chartier et al.'s, 2020, replication of Albarracín et al.'s, 2008, Experiment 5; $p = .012$ for Baranski et al.'s, 2020 replication of Shnabel & Nadler, 2008). It is possible that the expert feedback did reliably improve the effect size for the replication of Shnabel and Nadler (2008), but given the number of tests, it is also plausible that this difference occurred by chance. Therefore, even the most promising examples of formal peer review increasing replicability fail to provide reliable support.

We also examined the cumulative evidence for each of the 10 findings. Figure 2 shows the evidence from

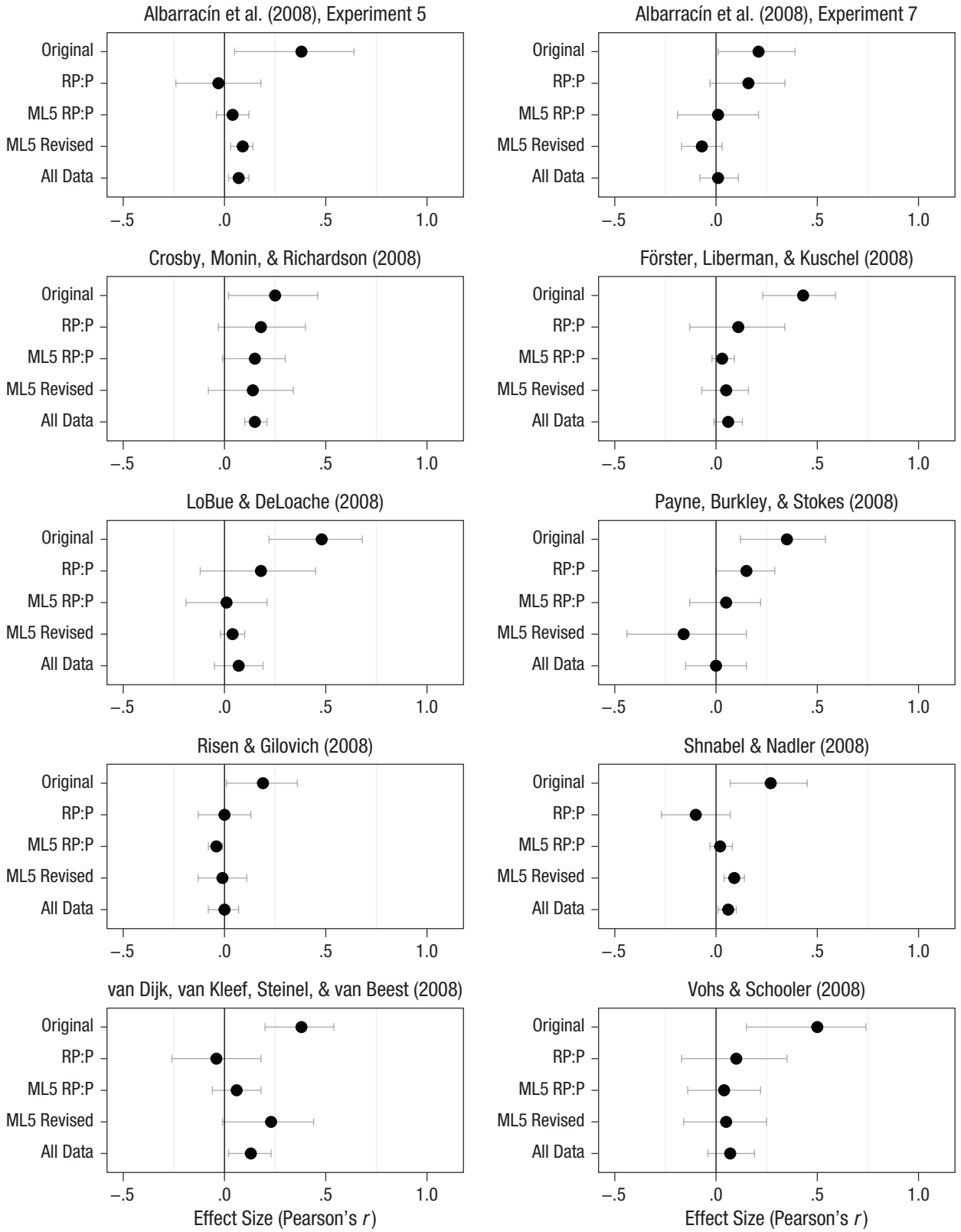

Fig. 2. Effect sizes from the 10 original studies and their replications in the Reproducibility Project: Psychology (RP:P; Open Science Collaboration, 2015) and the Many Labs 5 (ML5) protocols. The "All Data" results are estimates from random-effects meta-analyses including the original studies and their replications. Error bars represent 95% confidence intervals.

each original study, its RP:P replication, and both protocols in the current investigation, as well as the evidence combined from all four sources. The combined evidence provides the highest powered test to detect a small effect and the most precise estimate. For 4 of the 10 studies, the combined evidence indicated a statistically significant effect, though the effect sizes (median $r = .10$) were much smaller than the original reports' effect sizes (median $r = .37$), and all highest bounds of the 95% confidence intervals were below .25 (most were far below).

Exploratory analyses: additional measures of replicability

In exploratory analyses, we considered several other measures of replicability that directly assessed (a) statistical consistency between the replications and the original studies and (b) the strength of evidence provided by the replications for the scientific effects under investigation (Mathur & VanderWeele, 2020). These analyses also accounted for potential heterogeneity in the replications and for the sample sizes 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 was consistent with the replications. P_{orig} thus assesses whether the effect sizes obtained in the replications were similar to those of the original study; small values of P_{orig} indicate less similarity and larger values indicate 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 effect in 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 agree in direction with the original effect, but also are meaningfully strong by two different criteria (i.e., $r > .10$ or $r > .20$). These metrics together assess whether replications provide stand-alone evidence for the scientific hypothesis, regardless of the estimate of the original study itself.

For each original study, we conducted these analyses for three subsets of replications: (a) all Many Labs 5 replications, regardless of protocol; (b) replications using the Many Labs 5 RP:P protocol; and (c) replications using the Many Labs 5 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 the aggregated statistics (Mathur & VanderWeele, 2020; see our Supplemental Material for general methodological information about this approach). For replication subsets that had a heterogeneity estimate of exactly 0 or that 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 values of P_{orig}; the mean percentages of effects stronger than $r = 0$, .1, and .2; the probability of significance agreement; and the aggregate effect sizes with their p values. Despite our close standardization of protocols across sites, 40% of the replication sets had heterogeneity estimates greater than 0 (see Table 4), which highlights 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 replication studies using the revised and the RP:P protocols, 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 each original study and its replication studies had come from the same distribution. Of the replication studies using the revised and RP:P protocols, 50% and 80%, respectively, provided fairly strong evidence for inconsistency with the original study ($P_{orig} \leq .05$), and 20% and 30%, respectively, provided strong evidence for inconsistency ($P_{orig} < .01$). Thus, results for both the revised and the RP:P protocols often suggested statistical inconsistency with the original study, even after accounting for effect heterogeneity and other sources of statistical variability.[7] However, heuristically, evidence for inconsistency might have been somewhat less pronounced in the replication studies using the revised protocol rather than the RP:P protocol.

Table 4. Metrics of Replication Success by Study and Protocol Version

Original study and set of replications	Number of studies (*k*)	Estimate (*r*)	*p* value	τ	P_{orig}	Probability of significance agreement	Estimated percentage of population effects		
							Above *r* = 0	Above *r* = .1	Above *r* = .2
Albarracín et al. (2008), Experiment 5									
All replications	9	.07 [.01, .12]	.023	0	.06	.98	100	0	0
RP:P protocol	1	.04 [−.04, .12]	.34	0	.05	.96	100	0	0
Revised protocol	8	.09 [.03, .14]	.006	0	.08	.98	100	0	0
Albarracín et al. (2008), Experiment 7									
All replications	14	−.02 [−.11, .07]	.65	.10	.12	.77	50 [0, 71]	21 [0, 64]	0
RP:P protocol	7	.01 [−.16, .18]	.87	.13	.25	.65	57 [0, 86]	29 [0, 57]	14 [0, 81]
Revised protocol	7	−.06 [−.17, .05]	.19	.06	.03	.84	14 [0, 86]	0	0
Crosby, Monin, & Richardson (2008)									
All replications	6	.14 [.07, .21]	.004	0	.62	.82	100	100	0
RP:P protocol	3	.15 [−.01, .30]	.06	0	.64	.80	100	100	0
Revised protocol	3	.14 [−.09, .35]	.12	0	.61	.75	100	100	0
Förster, Liberman, & Kuschel (2008)									
All replications	16	.04 [−.01, .09]	.10	0	< .001	1	100	0	0
RP:P protocol	8	.03 [−.02, .08]	.18	0	< .001	1	100	0	0
Revised protocol	8	.04 [−.06, .15]	.36	.07	.004	.99	75 [0, 100]	12 [0, 100]	0
LoBue & DeLoache (2008)									
All replications	8	.02 [−.06, .11]	.50	0	.001	1	100	0	0
RP:P protocol	4	.01 [−.22, .24]	.89	.05	.003	.99	50 [0, 100]	0	0
Revised protocol	4	.04 [−.03, .10]	.16	0	.001	1	100	0	0
Payne, Burkley, & Stokes (2008)									
All replications	8	−.06 [−.21, .09]	.40	.16	.04	.82	38 [0, 62]	25 [0, 50]	0
RP:P protocol	4	.05 [−.13, .22]	.46	.07	.03	.94	75 [0, 100]	50 [0, 100]	0
Revised protocol	4	−.16 [−.44, .15]	.20	.18	.03	.72	25 [0, 50]	0	0
Risen & Gilovich (2008)									
All replications	5	−.04 [−.14, .07]	.20	0	.02	.96	0	0	0
RP:P protocol	1	−.04 [−.08, −.01]	.02	0	.02	.95	0	0	0
Revised protocol	4	−.01 [−.18, .16]	.87	.01	.06	.83	0	0	0
Shnabel & Nadler (2008)									
All replications	16	.05 [.02, .09]	.009	0	.04	.99	100	0	0
RP:P protocol	8	.02 [−.03, .08]	.38	0	.02	.98	100	0	0
Revised protocol	8	.09 [.03, .14]	.008	0	.08	.98	100	0	0
van Dijk, van Kleef, Steinel, & van Beest (2008)									
All replications	10	.10 [−.01, .20]	.07	.01	.006	1	100	40 [0, 100]	0
RP:P protocol	6	.06 [−.06, .19]	.24	0	.002	1	100	0	0
Revised protocol	4	.23 [−.03, .45]	.07	0	.19	.97	100	100	100
Vohs & Schooler (2008)									
All replications	9	.04 [−.06, .15]	.37	.06	.01	.98	78	22 [0, 100]	0
RP:P protocol	4	.04 [−.15, .23]	.55	0	.01	.98	100	0	0
Revised protocol	5	.05 [−.16, .25]	.55	.11	.03	.94	80	20 [0, 100]	0

Note: The estimates (Pearson's *r* scale) and *p* values are from meta-analyses; τ is the meta-analytic estimate (Fisher's *z* scale) of the standard deviation of the population effects in the replications. P_{orig} is the probability that the original study's estimate would have been as extreme as actually observed if the original study was consistent with the replication studies. Probability of significance agreement is the probability that the meta-analytic estimate in the replication studies would be statistically significant and would agree in direction with the estimate in the original study if the original study and the replications were consistent. Values in brackets are 95% confidence intervals; when the confidence intervals could not be statistically estimated for the percentage metrics, they are omitted. RP:P = Reproducibility Project: Psychology.

Regarding the strength of evidence for the scientific hypotheses, for the replication studies using the revised protocols, on average, only 50% of population effects agreed in direction with the effects in the original studies (as expected if the average effect size were exactly zero), 20% were above a modest effect size of .10, and 10% were above an effect size of .20. For the replication studies using the RP:P protocols, on average, 60% of effects agreed in direction with the effects in the original studies, 10% were above an effects size of .10, and 0% were above and effect size of .20. These results suggest that even after accounting for heterogeneity, the large majority of population effects were negligibly small regardless of protocol version.[8] Thus, for both the revised and the RP:P protocols, the population effects did not reliably support the scientific hypotheses even when we used the generous criterion of considering all effects that agreed in direction with the effects in 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 the extent to which prediction markets and a survey 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 reported are based on the participants who made at least one trade.[9]

The survey and prediction markets produced collective peer estimates of the probability of success for each replication protocol. The mean predicted probability of a statistically significant replication effect 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: $p = .232$, $n = 10$). Thus, participants expected about 3 of 10 studies using each protocol type to replicate the original effects. As reported on the survey, participants believed, on average, that the likelihood of 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: $p = .002$, $n = 10$).[10]

The relationship between peer beliefs about replication success and replication outcomes (i.e., whether or not a significant replication effect was observed) is shown separately in Figure 3 for prediction-market beliefs (Fig. 3a) and survey beliefs (Fig. 3b). Both the prediction-market beliefs ($r = .07$, $p = .780$, $n = 20$) and the survey beliefs ($r = -.14$, $p = .544$, $n = 20$) were weakly 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 predictions and the 20 prediction-market predictions as independent observations, which may not hold because the predictions might have been correlated within each study. Pooling beliefs across protocols so that we had just 10 observations in each analysis yielded a point-biserial correlation of −.02 ($p = .956$) between the prediction-market beliefs and replication outcomes, a correlation of −.09 ($p = .812$) between the survey beliefs and the replication outcomes, and a correlation of .707 ($p = .022$) between the prediction-market beliefs and the survey beliefs.

Discussion

We tested whether revising protocols on the basis of formal peer review by experts could improve replication success for a sample of studies that had mostly failed to replicate original findings 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, generally, the revised protocols elicited effect sizes very similar to those of the replication protocols based on the RP:P. Neither of our primary analysis strategies led to rejection of the null hypothesis that formal peer review has no effect on replicability, and the estimates of the effect of protocol were very small, with very narrow confidence intervals ($\Delta r = .002$, 95% CI = [−.04, .04]; $\Delta r = .014$, 95% CI = [−.02, .05]). Analysis of the data from both the revised and the RP:P protocols provided evidence for statistical inconsistency with the original studies even across the varied contexts in which the 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, the exploratory analyses based on the original effect sizes and new samples indicated that the average expected percentages of significant results among the revised and RP:P replications were 90% and 92% (i.e., 9 of 10 replications; see Table 4), respectively (Mathur & VanderWeele, 2020). However, even focusing on the two significant results does not provide good evidence that peer review strengthens replication effect sizes. Just one of these two replication sets 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 effect sizes.

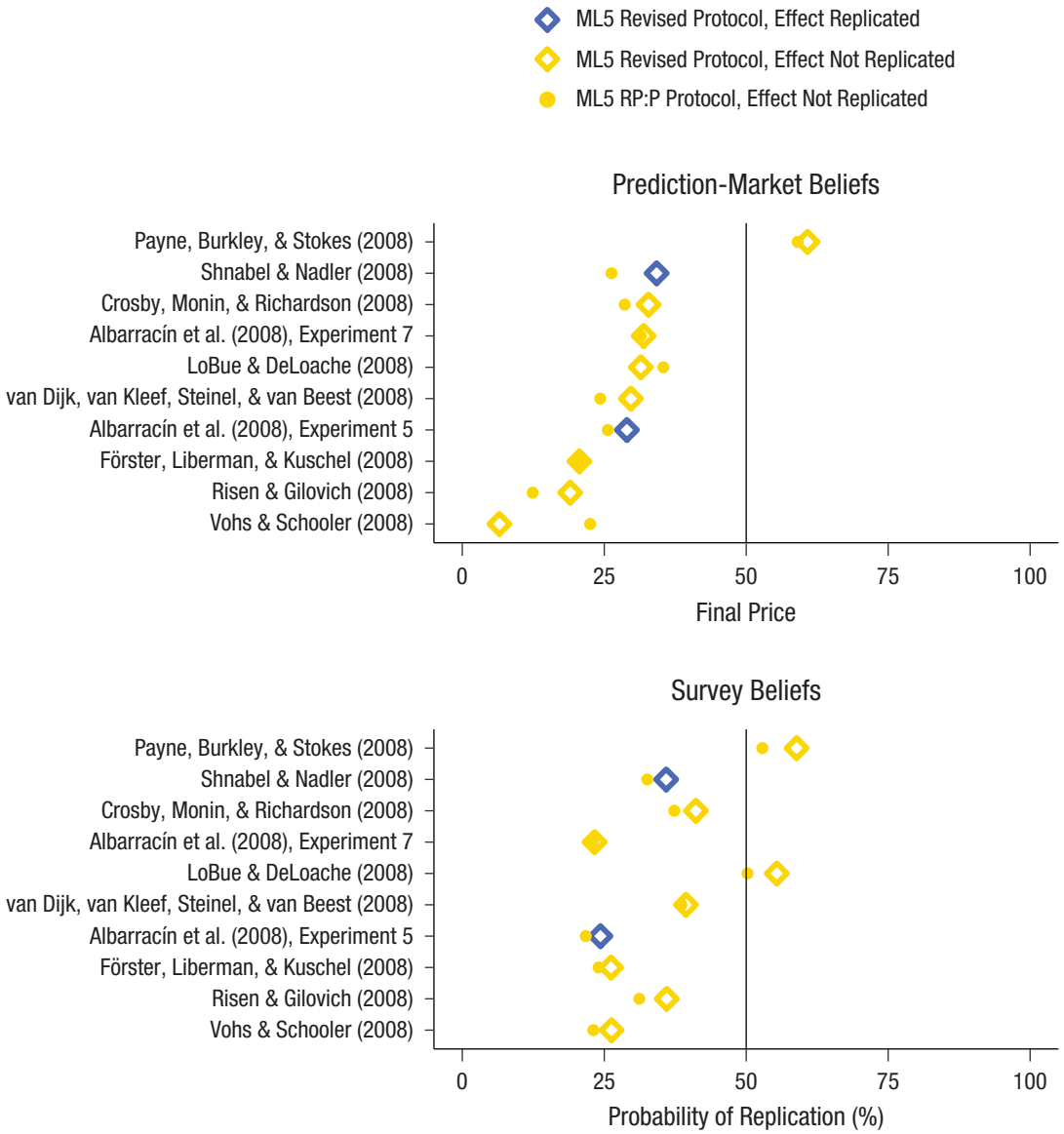

Fig. 3. Peer beliefs about replication outcomes. The graph in (a) shows the final price of each replication in the prediction market. The graph in (b) shows the rated probability that each study's result would be replicated (survey beliefs). In both panels, the studies are listed in descending order of the prediction-market prices for the revised protocols. ML5 = Many Labs 5; RP:P = Reproducibility Project: Psychology.

Finally, in an analysis of the cumulative data from the original, RP:P, and present studies, four of the effects were significant and in the same direction as the original effects, albeit with very small effect sizes. None exceeded an r value of .15, even though the original effect sizes had a median of .37 and a range of .19 to .50. All were quite precisely estimated, and the upper bounds of their 95% confidence intervals were .23 or less. Of the 111 replication effect sizes from the RP:P and this investigation, only 4 were as large as or larger than the effect size of the corresponding original finding (see Fig. 4). Indeed, exploratory analyses suggested that 50% and 80% of the replication studies using the revised and RP:P protocols, 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, the original effect sizes were extreme compared with all the effect sizes obtained in the attempts to reproduce them.

Conducting formal peer review did not increase observed effect sizes for the replication efforts, on average. We observed some evidence consistent with a few original findings, but the effect sizes were sharply lower

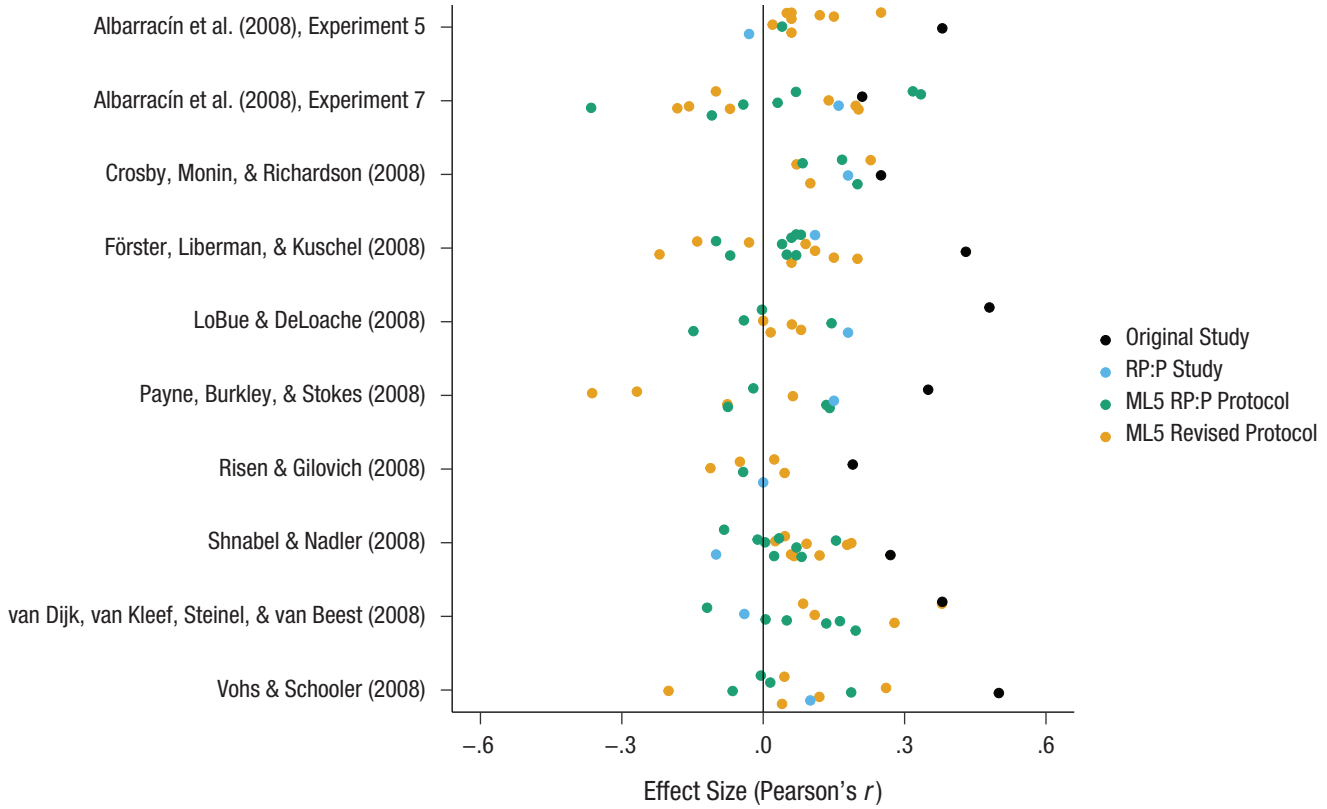

Fig. 4. Effect sizes from individual sites. Results are shown for each original study, its replication in the Reproducibility Project: Psychology (RP:P), and its Many Labs 5 (ML5) replications.

regardless of protocol. 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 the findings obtained in the replication efforts.

Finally, neither prediction markets nor the survey 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 in the current project could be that the small sizes of the samples of traders and studies produced uncertain estimates (past markets have involved 40–80 traders and 20–30 studies; Dreber et al., 2015; Forsell et al., 2019). Also, a floor effect may have occurred in that the replication effect sizes were generally much smaller than those of the original studies and provided 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 replicated the original findings. The present evidence provides mixed support at best for Gilbert et al.'s speculation.

The most optimistic conclusion can be drawn by focusing on the criterion of achieving statistical significance ($p < .05$). From that perspective, the rate of successful replication went from 0 out of 10 original studies with the RP:P protocol to 2 out of 10 with the revised protocol. It is easy for the optimist to conclude descriptively that adding peer review in advance and increasing power substantially increased replicability of the findings.

The most pessimistic conclusion is that even with extremely high power, the formal analyses did not support the hypothesis that peer review would increase replicability on average. Even among the significant results, only one of the two provided evidence consistent with that hypothesis. Moreover, 3 of the 10 revised protocols yielded effects in the direction opposite the direction of the original effects, despite high power and peer review. And perhaps most critically, effect sizes were dramatically smaller in these optimized replications compared with the original studies. The median effect size was .37 for the original findings, .11 for the RP:P, .04 for

the RP:P protocols in the current project, and .05 for the revised protocols in the current project. On average, the original studies would have had 22% power to detect effects of the magnitude produced by the corresponding revised protocols (excluding revised protocols that produced negative effect sizes). It is easy for the pessimist to conclude descriptively 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 the 10 findings. Figure 2 shows that, with all data combined, 4 of the 10 effects were statistically significant ($p < .05$), and all 10 effect sizes were quite precisely estimated and small (median $r = .07$; range = 0–.15). All 10 of the meta-analytic results are much smaller than the original findings (median $r = .37$; range = .19–.50). As data are accumulated, reliable results should be associated with p values approaching zero rather than remaining close to .05, which would indicate weak evidence (Benjamin et al., 2017). However, even with the data from the original studies retained, the 4 significant meta-analytic results do not have uniformly very small p-values approaching zero (Albarracín et al., 2008, Experiment 5: $p = .014$; Crosby, Monin, & Richardson, 2008: $p = .0004$; Shnabel & Nadler, 2008: $p = .015$; van Dijk, van Kleef, Steinel, & van Beest, 2008: $p = .023$). The most encouraging individual finding for demonstrating replicability is that for Crosby et al. (2008). None of the replication studies for Crosby et al. achieved statistical significance on their own, but the cumulative evidence supports the original finding, albeit with a reduced effect size. Notably, our results for this finding simultaneously showed no evidence of improved replicability based on peer review (the revised protocol elicited an effect size 44% weaker than that in the original study). The most parsimonious explanation for the observed data may be that the effect is weaker than indicated by the original study and not moderated by the factors that differed between the protocols.

In summary, some of the original findings may be replicable, and all the 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 will yield more significant effects and larger effect sizes (see Box 1 for potential future directions based on individual studies). The study for which the evidence for improvement through expert review was strongest (Shnabel & Nadler, 2008) provides a suggestive direction for such refinements. The primary revisions to the protocol for that study involved extensive tailoring and pilot-testing of study materials for new populations. However, this was not the only study whose revisions included this process, and it is possible that the apparent benefits of the revisions occurred by chance. Across all studies, the original findings were statistically anomalous compared with the replication findings, and the prediction markets, reviewers, and replication teams could not predict which findings would persist with some supporting evidence.

For those findings whose replicability did not improve through expert review, the present understanding of the conditions needed for replicating the effect is not sufficient. This minimally suggests that theoretical revisions are needed in order to understand the boundary conditions for observing the effect, and maximally suggests that the original result may actually be a false positive. In the latter case, it is possible that no amount of expertise could have produced a replication of the original 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 conditions necessary for replicability (Nosek & Errington, 2020).

Constraints on generality

There are two primary and related constraints on the generality of our conclusions regarding the role of expertise in peer review: the selection of studies investigated and statistical power. The original studies investigated in this project were selected because there was reason, a priori, to suspect that they could be improved through peer review. If the labeling of these studies as nonendorsed 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 nonendorsement from original authors might have selected for relatively unreliable 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 our findings will generalize is unknown. It is possible that our findings are unique to this sample of studies, or to psychology studies that are conducted in good faith but fail to be endorsed by original authors, as in the RP:P (Open Science Collaboration, 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

Box 1. Case Studies for Generating Hypotheses About Expertise

In the aggregate, our project indicated that expert peer review had little impact on improving replicability across the 10 original 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 differences between protocols). Consider the following examples:

- Albarracín et al. (2008): Two of our included studies came from Albarracín et al. (2008). These two studies yielded evidence that instilling action or inaction goals influences subsequent motor and cognitive output (Experiment 5: $r = .38$; Experiment 7: $r = .21$). In the Reproducibility Project: Psychology (RP:P), neither finding was replicated according to the statistical-significance criterion, but the effect size for the replication of Experiment 7 ($r = .16$) was close to the original. Experiment 5's replication elicited a small effect size in the direction opposite that of the original ($r = -.03$). The present replications likewise elicited small effect sizes, but with an interesting pattern. For Experiment 5, expert review was descriptively and not significantly ($p = .601$) associated with a larger effect size (RP:P protocol: $r = .04$; revised protocol: $r = .09$). For Experiment 7, expert review was descriptively and not significantly ($p = .150$) associated with an effect size in the wrong direction (RP:P protocol: $r = .02$; revised protocol: $r = -.07$). If these patterns are not just statistical noise, they signal an occasion for pursuing a perspectivist approach to understanding the role of expertise in replicability (McGuire, 2004): Under what conditions does expertise improve versus reduce replicability?
- Payne, Burkley, and Stokes (2008): These authors observed that implicit and explicit race attitudes were less strongly correlated when participants were told to respond without bias than 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 effect, but it was smaller than in the original study ($r = .15$); in contrast, the higher-powered replications with the RP:P ($r = .05$) and revised ($r = -.16$) protocols did not elicit significant effects. In fact, the revised protocol's effect size was in the wrong direction and was significantly different from the RP:P protocol's effect size ($p = .002$). Most provocatively, this pattern directly opposes our hypothesis that formal peer review can improve replicability. We suspect that the effect in question is weaker than originally observed, that the effect is nonexistent (i.e., the original was a false positive), or that the social context for observing the effect has changed.
- Shnabel and Nadler (2008): These authors 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 in the current project both elicited substantially weaker effect sizes, but the effect size of the revised protocol was slightly larger than that of the RP:P protocol ($p = .012$). Among the replication results for all 10 original studies, this pattern is most consistent with the hypothesis that expert review improves replicability. Even so, the results for the revised protocol provided an overall point estimate that was 67% smaller than the estimate in the original study. The fact that an effect of protocol was found for just 1 of the 10 original studies does increase the plausibility that this effect occurred by chance. Nevertheless, if the difference is replicable, then these protocols might help in studying 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., 2020, this issue), and yet the outcomes on the dependent variable landed on opposing sides of the statistical-significance criterion (ps = .004, .350).
- van Dijk, van Kleef, Steinel, and van Beest (2008): These authors observed that individuals made more generous offers in negotiations with happy negotiation partners compared with 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 ($r = .06$), but the difference between protocols was not significant ($p = .315$). However, if the difference between protocols 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 studied phenomenon whether experts guided the protocol development or not (Klein et al., 2019).

boundary conditions on the effects of expertise in that some expert-guided revisions to research designs will improve replicability under some conditions. It is unknown whether the conditions under which expert-guided revisions improve replicability 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 sufficiently diverse conditions for us to observe an effect of peer review. Therefore, 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 replication of previous results may be tempting given the very small effect of expert peer review on replication effect sizes that we observed. However, that interpretation is unwarranted. The present study was 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 original findings in a prior replication project, the RP:P. Those original replication protocols were developed in a structured process with original materials and preregistration, the replication researchers had sufficient self-identified expertise to design and conduct the replications, and the designs received informal review by an internal review process and by original authors when they were willing to provide it. This informal review did not preclude the possibility of errors, but using RP:P protocols meant that the control condition already involved substantial effort and expertise aimed at conducting a faithful replication. Whether that effort and expertise was sufficient was the open question. The intervention we tested as a potential means of improving 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 those studies in which researchers have already made good efforts to conduct high-quality replications, such as the systematic replication studies populating social-behavioral sciences over the past 10 years.

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 the targeted findings in presumably ideal circumstances is that those findings were false positives. If so, then this study does not offer a test of the effectiveness of peer review in improving the quality of study methodology. A finding must be replicable under some conditions to test whether different interventions influence its replicability. We did not observe any conditions under which several of the original findings were replicable (see also Klein et al., 2019).

There may be conditions under which these findings 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 conducted the original research. It is possible, despite the presumed expertise of the reviewers, that they lacked knowledge of what would make the findings 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 effects 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 because of their experience with instrumentation and protocols. The meaningful question of the role of expertise in improving replicability concerns situations in which 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.

Transparency

Action Editor: Daniel J. Simons
Editor: Daniel J. Simons
Author Contributions

C. R. Ebersole and B. A. Nosek conceived the project and drafted the report. M. B. Mathur and C. R. Ebersole designed the analysis plan and analyzed the aggregate data. C. R. Ebersole, C. R. Chartier, J. K. Hartshorne, H. IJzerman, I. Ropovik, M. B. Mathur, L. B. Lazarević, H. Rabagliati, M. Corley, E. Baranski, D.-J. Bart-Plange, K. S. Corker, and N. R. Buttrick served as team leaders for the sets of replications. D. Viganola, C. R. Ebersole, Y. Chen, T. Pfeiffer, A. Dreber, M. Johannesson, and B. A. Nosek designed and analyzed the survey and prediction markets to elicit peer beliefs. All the authors except B. A. Nosek collected the data. All the authors revised and approved the submitted

manuscript with two exceptions; sadly, S. Pessers and B. Petrović passed away before the manuscript was finalized.

Declaration of Conflicting Interests

B. A. Nosek is Executive Director of the nonprofit Center for Open Science, which has a mission to increase openness, integrity, and reproducibility of research. The author(s) declared that there were no other potential conflicts of interest with respect to the authorship or the publication of this article.

Funding

The authors thank the following sources of funding: the French National Research Agency (ANR-15-IDEX-02; H. IJzerman), the Netherlands Organization for Scientific Research (NWO; 016.145.049; H. IJzerman), the National Institute on Alcohol Abuse and Alcoholism (F31AA024358; M. H. Bernstein), the Social Sciences and Humanities Research Council of Canada (149084; M. Inzlicht), the Economic and Social Research Council (United Kingdom; ES/L01064X/1, H. Rabagliati), the John Templeton Foundation (C. R. Ebersole and B. A. Nosek), the Templeton World Charity Foundation (B. A. Nosek), and the Templeton Religion Trust (B. A. Nosek).

Open Practices

Open Data: https://osf.io/7a6rd/
Open Materials: not applicable
Preregistration: https://osf.io/nkmc4/
All data and analysis scripts have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/7a6rd/. The design and analysis plans were preregistered at the Open Science Framework and can be accessed at https://osf.io/nkmc4/. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/2515245920958687. This article has received badges for Open Data and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.



ORCID iD

Charles R. Ebersole https://orcid.org/0000-0002-8607-2579

Acknowledgments

The authors thank the many original authors and experts who provided extensive feedback throughout the many stages of the project.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/2515245920958687

Notes

1. The RP:P included 100 replications; however, 3 of the original studies showed null results.
2. There has been some confusion over the procedure for labeling endorsement of RP:P studies (e.g., Gilbert et al., 2016). Assessments of original authors' endorsement were made by replication teams prior to conducting the replications. They assessed what they believed the authors' endorsement to be, on the basis of whether or not the replication design had addressed any concerns raised by the original authors.
3. Correspondence from the RP:P (Open Science Collaboration, 2015) was accessed from that project's OSF page (osf.io/ezcuj/).
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. (2020, this issue).
5. The results we report focus on meta-analytic outcomes across the studies. Most of the individual reports of the 10 replication teams tend to use mixed-effects models to gauge statistical significance. As a result, the statistical significance of each protocol in a given study may differ here from what is reported in the individual report.
6. 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.
7. 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 replications using the revised and the RP:P protocols, respectively. Of the studies using the revised and RP:P protocols, 20% and 86%, respectively, had P_{orig} values less than .05, and 20% and 29%, respectively, had P_{orig} values less than .01.
8. In sensitivity analyses as described in the previous note, we estimated that 100%, 40%, and 20% of effects in the replications using the revised protocols were stronger than $r = 0$, $r = .1$, and $r = .2$, respectively. We estimated that 86%, 14%, and 0% of effects in the replications using the RP:P protocols were stronger than these thresholds.
9. Many Labs 5 contributors were not allowed to make predictions on their studies, and their survey answers about those studies were not used.
10. 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%)?"

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*, 510–523.

Anderson, C. J., Bahník, Š., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., . . . Zuni, K. (2016). Response to Comment on "Estimating the reproducibility of psychological science." *Science*, *351*, 1037.

Andrews, I., & Kasy, M. (2019). Identification of and correction for publication bias. *American Economic Review*, *109*, 2766–2794.

Baranski, E., Baskin, E., Coary, S., Ebersole, C. R., Krueger, L. E., Lazarević, L. B., . . . Žeželj, I. (2020). Many Labs 5: Registered Replication of Shnabel and Nadler (2008), Study 4. *Advances in Methods and Practices in Psychological Science*, *3*, 405–417.

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 Behaviour*, *2*, 6–10. doi:10.1038/s41562-017-0189-z

Buttrick, N. R., Aczel, B., Aeschbach, L. F., Bakos, B. E., Brühlmann, F., Claypool, H. M., . . . Wood, M. J. (2020). Many Labs 5: Registered Replication of Vohs and Schooler (2008), Experiment 1. *Advances in Methods and Practices in Psychological Science*, *3*, 429–438.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., . . . Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*, 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*, 637–644.

Center for Open Science. (n.d.). *Registered Reports: Peer review before results are known to align scientific values and practices*. Retrieved from http://cos.io/rr/

Chambers, C. D. (2013). *Registered Reports*: A new publishing initiative at *Cortex*. *Cortex*, *49*, 609–610.

Chartier, C. R., Arnal, J. D., Arrow, H., Bloxsom, N. G., Bonfiglio, D. B. V., Brumbaugh, C. C., . . . Tocco, C. (2020). Many Labs 5: Registered Replication of Albarracín et al. (2008), Experiment 5. *Advances in Methods and Practices in Psychological Science*, *3*, 332–339.

Chen, S.-C., Szabelska, A., Chartier, C. R., Kekecs, Z., Lynott, D., Bernabeu, P., . . . Schmidt, K. (2018). *Investigating object orientation effects across 14 languages*. doi:10.31234/osf.io/t2pjv

Corker, K. S., Arnal, J. D., Bonfiglio, D. B. V., Curran, P. G., Chartier, C. R., Chopik, W. J., . . . Wiggins, B. J. (2020). Many Labs 5: Registered Replication of Albarracín et al. (2008), Experiment 7. *Advances in Methods and Practices in Psychological Science*, *3*, 340–352.

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*. Advance online publication. doi:10.1007/s13164-018-0400-9

Crosby, J. R., Monin, B., & Richardson, D. (2008). Where do we look during potentially offensive behavior? *Psychological Science*, *19*, 226–228. doi:10.1111/j.1467-9280.2008.02072.x

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

Ebersole, C. R., Andrighetto, L., Casini, E., Chiorri, C., Dalla Rosa, A., Domaneschi, F., . . . Vianello, M. (2020). Many Labs 5: Registered Replication of Payne, Burkley, and Stokes (2008), Study 4. *Advances in Methods and Practices in Psychological Science*, *3*, 387–393.

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.

Forsell, E., Viganola, D., Pfeiffer, T., Alemberg, J., Wilson, B., Chen, Y., . . . Dreber, A. (2019). Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology*, *75*(Part A), Article 102199. doi:10.1016/j.joep.2018.10.009

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*, 579–599.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*, 460–466.

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

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

IJzerman, H., Ropovik, I., Ebersole, C. R., Tidwell, N. D., Markiewicz, Ł., Souza de Lima, T. J., . . . Day, C. R. (2020). Many Labs 5: Registered Replication of Förster, Liberman, and Kuschel's (2008) Study 1. *Advances in Methods and Practices in Psychological Science*, *3*, 366–376.

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

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C. A., Nosek, B. A., Chartier, C. R., . . . Ratliff, K. A. (2019). *Many Labs 4: Failure to replicate mortality salience effect with and without original author involvement*. doi:10.31234/osf.io/vef2c

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*, 142–152.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., . . . Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*, 443–490. doi:10.1177/2515245918810225

Lazarević, L. B., Purić, D., Žeželj, I., Belopavlović, R., Bodroža, B., Čolić, M. V., . . . Stojilović, D. (2020). Many Labs 5: Registered Replication of LoBue and DeLoache (2008). *Advances in Methods and Practices in Psychological Science*, *3*, 377–386.

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*, 284–289. doi:10.1111/j.1467-9280.2008.02081.x

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*, 537–542.

Manski, C. F. (2006). Interpreting the predictions of prediction markets. *Economics Letters*, *91*, 425-429.

Mathur, M. B., Bart-Plange, D.-J., Aczel, B., Bernstein, M. H., Ciunci, A. M., Ebersole, C. R., . . . Frank, M. C. (2020). Many Labs 5: Registered multisite replication of the tempting-fate effects in Risen and Gilovich (2008). *Advances in Methods and Practices in Psychological Science*, *3*, 394–404.

Mathur, M. B., & VanderWeele, T. J. (2020). New statistical metrics for multisite replication projects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. Advance online publication. doi: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*, 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*, 657–664.

Nosek, B. A., & Errington, T. M. (2017). Reproducibility in cancer biology: Making sense of replications. *Elife*, *6*, Article e23383. doi:10.7554/eLife.23383

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

Nosek, B. A., & Gilbert, E. A. (2016). Let's not mischaracterize the replication studies. *Retraction Watch*. Retrieved from https://retractionwatch.com/2016/03/07/lets-not-mischaracterize-replication-studies-authors/

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*, Article aac4716. doi:10.1126/science.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*, 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*, 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.

Rabagliati, H., Corley, M., Dering, B., Hancock, P. J. B., King, J. P. J., Levitan, C. A., . . . Millen, A. E. (2020). Many Labs 5: Registered Replication of Crosby, Monin, and Richardson (2008). *Advances in Methods and Practices in Psychological Science*, *3*, 353–365.

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

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 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*, 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*, 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*, 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*, 1123–1128.

Skorb, L., Aczel, B., Bakos, B. E., Feinberg, L., Hałasa, E., Kauff, M., . . . Hartshorne, J. K. (2020). Many Labs 5: Replication of van Dijk, van Kleef, Steinel, and van Beest (2008). *Advances in Methods and Practices in Psychological Science*, *3*, 418–428.

Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science*, *12*, 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*, 30–34.

Strack, F. (2016). Reflection on the smiling Registered Replication Report. *Perspectives on Psychological Science*, *11*, 929–930.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*, 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*, 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*, 49–54.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral & Brain Sciences*, *41*, Article E120. doi:10.1017/S0140525X17001972