Check for updates

General Article

Advances in Methods and Practices in Psychological Science
2018, Vol. 1(3) 389–402
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2515245918787489
www.psychologicalscience.org/AMPPS

A Unified Framework to Quantify the Credibility of Scientific Findings

Etienne P. LeBel[1], Randy J. McCarthy[2], Brian D. Earp[3,4], Malte Elson[5], and Wolf Vanpaemel[6]

[1]Department of Psychology, University of Western Ontario; [2]Center for the Study of Family Violence and Sexual Assault, Northern Illinois University; [3]Department of Philosophy, Yale University; [4]Department of Psychology, Yale University; [5]Psychology of Human Technology Interaction Unit, Ruhr University Bochum; and [6]Quantitative Psychology and Individual Differences Unit, University of Leuven (KU Leuven)

Abstract

Societies invest in scientific studies to better understand the world and attempt to harness such improved understanding to address pressing societal problems. Published research, however, can be useful for theory or application only if it is credible. In science, a credible finding is one that has repeatedly survived risky falsification attempts. However, state-of-the-art meta-analytic approaches cannot determine the credibility of an effect because they do not account for the extent to which each included study has survived such attempted falsification. To overcome this problem, we outline a unified framework for estimating the credibility of published research by examining four fundamental falsifiability-related dimensions: (a) transparency of the methods and data, (b) reproducibility of the results when the same data-processing and analytic decisions are reapplied, (c) robustness of the results to different data-processing and analytic decisions, and (d) replicability of the effect. This framework includes a standardized workflow in which the degree to which a finding has survived scrutiny is quantified along these four facets of credibility. The framework is demonstrated by applying it to published replications in the psychology literature. Finally, we outline a Web implementation of the framework and conclude by encouraging the community of researchers to contribute to the development and crowdsourcing of this platform.

Keywords

research credibility, transparency, open science, analytic reproducibility, analytic robustness, replicability

Received 6/13/17; Revision accepted 6/6/18

Every year, societies spend billions of dollars to fund scientific research aimed at deepening understanding of the natural and social world. It is expected that some of the insights revealed by that research will lead to applications that address pressing social, medical, and other problems. Published research, however, can be useful for theory or applications only if it is *credible*. In science, a credible finding or hypothesis is one that has repeatedly survived high-quality, risky attempts at proving it wrong (Lakatos, 1970; Popper, 1959). The more such falsification attempts a finding survives, and the riskier those attempts are, the more credible a finding can be considered.

The currently dominant strategy to assess the credibility of an effect involves meta-analyzing all known studies on that effect (e.g., Cooper, Hedges, & Valentine, 2009). Such state-of-the-art meta-analytic approaches, however, cannot determine the true credibility of an effect because they do not account for the extent to which each included study has survived risky falsification attempts. For instance, the transparency, analytic credibility, and methodological similarity of meta-analyzed studies are not accounted for (even the standard methods used in Cochrane Reviews of medical research suffer from these limitations; Higgins, Lasserson, Chandler, Tovey, & Churchill, 2018). A credible finding must survive scrutiny

Corresponding Author:
Etienne P. LeBel, Department of Psychology, Social Science Centre, University of Western Ontario, 1151 Richmond St., London, Ontario, Canada N6A 3K7
E-mail: etienne.lebel@gmail.com

along four fundamental kinds of falsifiability-related dimensions:

- *Method and data transparency:* availability of design details, analytic choices, and underlying data);
- *Analytic reproducibility:* ability of reported results to be reproduced by repeating the same data processing and statistical analyses on the original data);
- *Analytic robustness:* robustness of results to different data-processing and data-analytic decisions); and
- *Effect replicability:* ability of the effect to be consistently observed in new samples, at a magnitude similar to that originally reported, when methodologies and conditions similar to those of the original study are used (see Appendix A at https://osf.io/gpu3a for more details regarding terminology).

If a finding withstands scrutiny along these four dimensions, such that independent researchers fail to identify fatal design flaws, data-processing or statistical errors or fragilities, or replicability issues, then an effect can be (temporarily) retained as not yet falsified and hence treated as credible.[1] The more intense the scrutiny along these four dimensions that a finding survives (i.e., the riskier the falsification attempts), the more one can be justified in treating it as credible (Popper, 1959).

Accordingly, to determine a finding's credibility, one must assess the degree to which it is transparent, reproducible, robust, and replicable. Quantifying these falsifiability-related properties, however, requires a systematic approach because they are interrelated. Information about one property may influence judgments about the other properties.

Currently, some initiatives do archive information about studies' analytic reproducibility, analytic robustness, and replications in new samples—for example, ReplicationWiki (http://replication.uni-goettingen.de/wiki/index.php/Main_Page) for economics, Harvard Dataverse (https://dataverse.harvard.edu/) for political science, PsychFileDrawer (http://psychfiledrawer.org/) for psychology, and Replications in Experimental Philosophy (http://experimental-philosophy.yale.edu/xphipage/Experimental%20Philosophy-Replications.html) for experimental philosophy. These projects, however, are limited by a lack of standardization, which prevents precise estimation of reproducibility, robustness, and replicability across studies and research fields. In the reproducibility and robustness archives, no standardized workflow is used to guide researchers on which reproducibility and robustness analyses to conduct, and no standardized scoring procedure is used to quantify the degree of reproducibility and robustness observed. In the replication archives, the degree of transparency and methodological similarity of replications are not assessed, which precludes the estimation of replicability within and across operationalizations of an effect. Finally, none of these platforms archive information pertinent to all four dimensions.

To overcome these limitations, we outline a single, coherent framework for gauging the credibility of published findings. Guided by sophisticated falsificationist principles (Lakatos, 1970; Popper, 1959), we propose a unique standardized workflow in which researchers quantify a finding's degree of transparency, reproducibility, robustness, and replicability, and we outline a Web implementation of this framework currently in development.

The Curation Framework

We propose a unified curation framework that can be used to systematically evaluate the credibility of empirical research by quantifying its transparency, reproducibility, robustness, and replicability. Currently, no such unified framework exists, but assessing the degree to which a finding has survived scrutiny along these four dimensions is crucial to comprehensively minimize all forms of publication and researcher biases. Further, these dimensions are inherently interrelated and thus should generally be assessed in a particular order[2]: Knowledge about certain aspects is either necessary for or influences evaluations of other aspects (e.g., insufficient transparency may prevent the estimation of reproducibility and replicability; lack of robustness may call into doubt the value of executing a replication when an expensive design or difficult-to-recruit population is required). Indeed, this framework is the only one in which the transparency, reproducibility, robustness, and replicability of a finding are evaluated within a harmonized system logically ordered to maximize research efficiency. In brief, these dimensions are incorporated as follows:

1. Transparency: The proposed framework supports assessment of transparency by curating published articles' compliance to the *basic-4* reporting standard (LeBel et al., 2013) as well as more comprehensive reporting standards (e.g., CONSORT—Schulz, Altman, & Moher, 2010; STROBE—Vandenbroucke et al., 2014). In addition, open-practice badges (open materials, open data, and preregistration) are curated and linked to the corresponding publicly accessible content (even if an article is published in a journal that does not yet offer badges).

2. Analytic reproducibility: The proposed framework uses a standardized workflow to allow independent evaluation of the analytic reproducibility of a study's primary substantive finding (i.e., its primary outcome or set of outcomes, defined by Hardwick et al., 2018, as what is emphasized in an article's abstract, figures, or tables) and includes a scoring procedure to quantify the degree of analytic reproducibility observed.
3. Analytic robustness: The proposed framework employs a standardized workflow to allow independent investigations of the analytic robustness of a study's primary substantive finding and includes a scoring procedure to quantify the degree of analytic robustness observed.
4. Effect replicability: The proposed framework addresses the problems of publication and researcher bias by uniquely incorporating a falsifiability-informed approach to organizing and evaluating replication studies within and across methods and populations. To support such evaluation, key characteristics of replication studies are curated. These characteristics include methodological similarity to the original study, differences from the original study's design, evidence provided regarding the plausibility of auxiliary hypotheses (e.g., integrity of instruments), and independence of the investigators. A novel meta-analytic and individual-study statistical approach is used to evaluate replication results in a nuanced manner.

Curation of transparency

The first, and most fundamental, credibility facet to consider is the degree to which a study's methodological details and data are transparently reported. When sufficient methodological details concerning how a study was conducted are not available, it is impossible to comprehensively identify flaws in the study or errors in the data, and it is impossible to conduct independent replications. Consequently, the substantive hypothesis tested in a study reported without sufficient transparency is not falsifiable; that is, it is nearly impossible to prove the hypothesis wrong if it is in fact false (Feynman, 1974). In contrast, a high level of transparency affords a relatively high degree of falsifiability (Popper, 1959), increasing the likelihood that a false hypothesis will be proven wrong.

Four different aspects of transparency should be considered for original and replication studies. In descending order of how fundamental they are to transparency, these aspects are (a) compliance with reporting standards for the study design used, (b) open (i.e., publicly available) materials, (c) preregistration information, and (d) open data.

Compliance with reporting standards. Reporting standards are crucial because they specify the precise methodological details that need to be reported given the specific kind of study design employed. When such information is transparently reported, researchers are in a position to identify flaws and confirm that rigorous methodology was indeed used. If such information is not reported, it is impossible to evaluate the rigor of a study.[3]

Prior to 2011, psychology journals did not mandate compliance with official reporting guidelines (though some researchers were advocating that this be done; see, e.g., Kashy, Donnellan, Ackerman, & Russell, 2009). Demonstrating how easy it is to provide "evidence" for a false conclusion with then-current reporting standards by intentionally or unintentionally exploiting design and analytic flexibility, Simmons, Nelson, and Simonsohn (2011) proposed a disclosure-based solution whereby authors are required to disclose five basic methodological details about how a study was conducted.

Inspired by Simmons, Nelson, and Simonsohn's (2012) subsequent 21-word solution, LeBel et al. (2013) then developed and popularized the basic-4 reporting standard through their grassroots initiative, PsychDisclosure.org. This initiative involved inviting 630 authors of recently published articles to disclose four methodological details that were not required to be reported but are crucial for accurate interpretation of published findings (i.e., excluded observations, all tested experimental conditions, all assessed outcome measures, and the rule for determining the sample size). About 50% of the contacted authors voluntarily disclosed this information, and the success of the initiative led psychology's flagship empirical journal, *Psychological Science* (Eich, 2014), and eventually several other journals, to require disclosure of these four methodological details when an article is submitted (LeBel & John, 2017).

Despite being an improvement over previous reporting standards in psychology, the basic-4 reporting standard still falls short of standards that have existed in the medical literature since the 1990s. (The CONSORT reporting guideline, e.g., specifies 25 methodological details that should be reported for any randomized controlled trial—or any experimental study; Begg et al., 1996; Schulz et al., 2010). As a starting point for our curation framework, we propose that, at a minimum, reporting of methodological details for the basic-4 categories should be curated. In the future, compliance with more thorough official reporting guidelines should be curated.

Open materials. Providing *open materials* means making all experimental materials and procedures required to

conduct a fair replication accessible in a public repository (Kidwell et al., 2016). This practice increases falsifiability of a tested hypothesis by substantially facilitating direct replications by independent researchers. It also increases falsifiability by allowing more thorough scrutiny of materials and procedures, which increases the likelihood that methodological shortcomings can be identified.

Preregistration information. Preregistration of a study's design and of analytic plans is crucial to transparency because it minimizes design and analytic flexibility that can be intentionally or unintentionally exploited (Nosek, Ebersole, DeHaven, & Mellor, 2018). Preregistration, whether done independently or through a registered-report format (Chambers, 2013), allows for more accurate adjustments for multiple analyses, and for clearer distinctions between confirmatory and exploratory analyses (assuming that the preregistered plan was sufficiently detailed and actually followed). Enhanced transparency in each of these respects also increases falsifiability.

Open data. Making data open is analogous to making materials open. It is the practice of making an article's raw (or transformed) original data accessible at a public repository. Such practice increases falsifiability because it allows independent researchers to scrutinize the integrity of the data (e.g., to confirm the number of participants and variables), which increases the likelihood of detecting data errors and internal data inconsistencies. If no serious errors are detected, then a researcher can be more confident regarding the reported results. Raw data are more transparent than transformed data and allow for even higher levels of falsifiability.

Summary. The proposed framework curates and organizes information on transparency, and links the open-practice badges to their respective content at the chosen public repository. This is done at the study level within articles, including those published in journals that do not yet award open-practice badges. Curating such information substantially increases the ease of finding it and, hence, increases falsifiability by making it easier for other researchers to detect any design or data errors. Indeed, the full value of increased transparency can be achieved only if such information is accessible and easy to find.

Curation of analytic reproducibility and robustness

Our proposed framework includes a standardized workflow for gauging the analytic reproducibility and analytic robustness of published findings. This workflow includes scoring procedures to quantify the degree to which a study's primary reported findings are analytically reproducible and analytically robust.

Analytic-reproducibility workflow. The proposed workflow specifies a standardized approach to guide independent researchers in verifying analytic reproducibility, that is, in determining whether the original primary substantive finding (as defined by Hardwicke et al., 2018) is again obtained when the original data-processing choices and statistical analyses are reapplied to the original (raw or transformed) data. In the proposed scoring procedure, the degree of analytic reproducibility is quantified by calculating the percentage of reproduced effect sizes (ESs) that are consistent with the corresponding originally reported ESs within a 10% margin of error (Hardwicke et al., 2018; see Appendix B at https://osf.io/gpu3a/ for more details).

If an independent researcher successfully confirms the analytic reproducibility of a study's reported primary finding, detecting no serious discrepancies (> 10%) between the reproduced and originally reported ESs, then the researcher can be more confident in the study's reported results. Under these conditions, it is justifiable to investigate the analytic robustness of the reported results. But if reproducibility cannot be assessed, because of insufficient description of data-processing or statistical choices, or if such an assessment yields discrepant results, then confidence in the reported results should be reduced. In such a situation, it is unclear whether the expenditure of time and resources to evaluate the analytic robustness of the study's reported results is justified.

Analytic-robustness workflow. The proposed workflow and scoring procedure for evaluating analytic robustness roughly parallel the workflow and scoring procedure for evaluating analytic reproducibility. The workflow is informed by Steegen, Tuerlinckx, Gelman, and Vanpaemel's (2016) multiverse analytic approach, in which one estimates the extent to which a study's conclusions are robust to reasonable alternative data-processing choices by examining the distribution of p values obtained for all combinations (a *multiverse*) of all such alternative data-processing choices. The workflow also is informed by Simonsohn, Simmons, and Nelson's (2015) specification-curve analysis, in which one estimates a study's primary effect using all reasonable combinations of alternative data-processing choices and statistical analyses (what Simonsohn et al. call specifications) and then conducts statistical tests to determine whether the set of such estimates is inconsistent with the null hypothesis.

The standardized workflow involves using all reasonable combinations of alternative data-processing choices and statistical analytic models to obtain a multiverse of ES estimates, with corresponding confidence intervals, for a study's primary substantive finding (as defined earlier). The degree of analytic robustness of a reported result is quantified by calculating the percentage of such

multiverse ES estimates that are consistent with the originally reported ES point estimate. When the reported result appears to be analytically robust (e.g., > 80% of the multiverse ES estimates are consistent with the original ES point estimate), it is justifiable to consider evaluating the replicability of the target substantive hypothesis (assuming the study's methodology has been reported sufficiently transparently, as described earlier). In contrast, if a reported result is not analytically robust (i.e., it is highly contingent on data-processing choices and analytic models), then evaluating its replicability may not be justified, depending on the resource costs or feasibility of conducting independent replications (e.g., it may not be justified to conduct a replication of a longitudinal study or a study involving a difficult-to-recruit population).

Curation of effect replicability

Our approach supports evaluation of effect replicability by specifying

- a flexible workflow in which replications are organized according to their distinct operationalization of a target effect (which also makes it possible to gauge the generalizability of an effect);
- curation of key characteristics of replication studies, including their methodological similarity to and differences from the original studies, the evidence they provide regarding the plausibility of auxiliary hypotheses, and the independence of the investigators; and
- a statistical approach in which meta-analysis and study-level analyses are used to evaluate replication results in a nuanced manner.

Such a falsifiability-informed and stringent approach minimizes the negative effects of all forms of publication and researcher biases, as we elaborate on in this section.

Flexible structure for researchers to organize replications. In the proposed framework, replications are organized according to their operationalization of an effect (see Appendix C at https://osf.io/gpu3a/ for a diagram). Replicability is gauged within a distinct operationalization, across replications that are sufficiently methodologically similar to the original study. The generalizability of an effect is evaluated by examining the degree to which it is replicable across distinct methodologies (i.e., different operationalizations of the independent and dependent variables) or distinct populations.

Curation of key characteristics of replication studies. The key characteristics that are curated for replication studies are (a) methodological similarity to the original study as determined using a principled replication taxonomy, (b) differences from the original design, (c) evidence of the plausibility of auxiliary hypotheses (e.g., integrity of instruments), and (d) investigator independence (see Appendix C at https://osf.io/gpu3a/ for more details regarding these characteristics). Note that transparency,[4] analytic reproducibility, and analytic robustness should also be examined and considered for replication studies, given that it is crucial to verify these characteristics for *all* studies.

Methodological similarity to the original study. To be eligible for inclusion in a collection of replication evidence, a replication study must employ a methodology that is sufficiently similar to the original study's methodology (Earp, in press). To guide the classification of replications according to their methodological similarity to an original study, we use the replication taxonomy depicted in Figure 1 (from LeBel, Berker, Campbell, & Loving, 2017), which is a simplified version of earlier taxonomies (Hendrick, 1991; Schmidt, 2009). In this taxonomy, replications range from a "highly similar" pole to a "highly dissimilar" pole (for more details and examples of the replication types along this continuum, see Appendix C at https://osf.io/gpu3a/).[5]

Each type of replication serves a different epistemological purpose (Zwaan, Etz, Lucas, & Donnellan, 2017). We consider only *direct replications* ("exact," "very close," and "close" replications) as sufficiently similar to an original study to be included in a collection of replication evidence. The reason for this is that only the results of these types of replications can in principle—across several replication attempts—falsify a hypothesis (assuming sound auxiliary hypotheses) and consequently cast doubt on the credibility of an effect (Earp & Trafimow, 2015; Meehl, 1967, 1978). By contrast, the major—and intentionally introduced— methodological differences of "far" and "very far" replications (i.e., generalizations) can never cast doubt on an originally reported effect (Doyen, Klein, Simons, & Cleeremans, 2014). This is because unsupportive evidence from such studies is ambiguous: It could be due to the falsity of the original hypothesis or to one or more of the changes in methodology in the replication attempt (Pashler & Harris, 2012). Hence, such studies can speak only to the generalizability of a presumably replicable effect.

In summary, only direct replications with methodology sufficiently similar to that of the original study, which naturally are constrained in design and analytic approach, can provide the sort of strict falsification attempt that

Replication Continuum

Highly Similar ← → Highly Dissimilar

	Direct Replication			Conceptual Replication	
Design Facet	Exact Replication (All facets under researcher control are the same)	Very Close Replication (Procedure or physical setting is different)	Close Replication (IV or DV stimuli are different)	Far Replication (IV or DV operationalization or population is different)	Very Far Replication (IV or DV constructs are different)
Effect, Hypothesis	Same	Same	Same	Same	Same
IV Construct	Same	Same	Same	Same	Different
DV Construct	Same	Same	Same	Same	Different
IV Operationalization	Same	Same	Same	Different	
DV Operationalization	Same	Same	Same	Different	
Population (e.g., age)	Same	Same	Same	Different	
IV Stimuli	Same	Same	Different		
DV Stimuli	Same	Same	Different		
Procedural Details	Same	Different			
Physical Setting	Same	Different			
Contextual Variables	Different				
⋮	⋮				

Fig. 1. Taxonomy for classifying a replication study's methodological similarity to an original study. "Same" indicates that the design facet in question is the same as in the original study, and "different" indicates that it is different. IV = independent variable; DV = dependent variable. "Population" refers to major population characteristics, such as age and whether the sample is drawn from the community or a special clinical population. Procedural details are minor experimental particulars (e.g., task instructions, font, font size). Contextual variables are design facets beyond a researcher's control (e.g., history, culture, language).

justifies increased confidence in a target hypothesis when the effect survives the falsification attempt. Such an approach contrasts sharply with the traditional meta-analytic approach, which cannot yield trustworthy conclusions because it combines incomparable studies of unknown methodological similarity and unknown levels of transparency, reproducibility, and robustness (for more details about inadequacies of traditional meta-analyses, see Appendix C at https://osf.io/gpu3a/).

Design differences. Design differences are any design characteristics, within or beyond the researcher's control, that differ from those of an original study. These are important to consider in order to arrive at an accurate interpretation of a replication result. Positive replication evidence shows that an effect is robust across the known design differences. When replication evidence is negative, such differences provide initial clues regarding potential boundary conditions of an effect.

Evidence regarding the plausibility of auxiliary hypotheses. A test of a substantive hypothesis rests on the assumption that several auxiliary hypotheses hold true (e.g., that the measurement instruments operated correctly; that participants understood the instructions and paid sufficient attention; Meehl, 1967). When researchers interpret study results, it is important for them to consider evidence that can help them gauge how plausible it is that such auxiliary hypotheses were sound (LeBel & Peters, 2011). Consequently, such information (also known as positive controls; Moery & Calin-Jageman, 2016) is a key study characteristic that is curated in our proposed framework. It is particularly important to consider this information when a replication study yields a null finding, to rule out more mundane explanations for the target effect not having been detected. For example, evidence of a successful manipulation check or detection of a known replicable effect (e.g., a semantic priming effect) helps rule out the possibility that a fatal experimenter error or data-processing error caused the observed null finding.

Investigator independence. Basic information about the degree of independence between the replication investigators and the researchers who conducted the original study is also key in interpreting replication results. Investigator independence is important to protect against confirmation and other biases (Earp & Trafimow, 2015; Rosenthal, 1991).

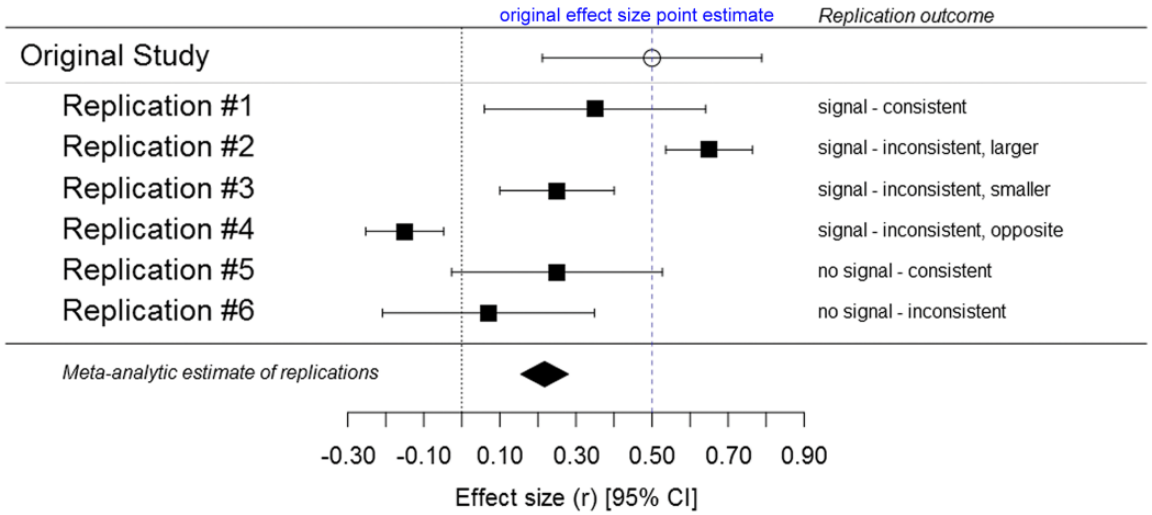

Fig. 2. Six hypothetical replication outcomes illustrating the three statistical aspects that should be considered when researchers interpret a replication result: (a) whether a signal was detected (i.e., whether the 95% confidence interval, or CI, represented here by the error bars, includes 0), (b) the consistency of the replication effect-size (ES) estimate with that observed in the original study (i.e., whether the replication's CI includes the original ES point estimate), and (c) the precision of the replication's ES estimate (i.e., the width of its CI relative to the CI in the original study). This figure is a reprint of Figure 1 in LeBel, Vanpaemel, Cheung, and Campbell (2018), which was published under a CC-By Attribution 4.0 International license.

Summary. Curating these four key characteristics of replication studies helps researchers evaluate the replicability of an effect in a nuanced fashion by allowing them to weight replications according to these dimensions (e.g., researchers can then give more weight to more transparently reported or analytically reproducible replications and give less weight to replications conducted by nonindependent researchers). Such information also allows researchers to quantitatively meta-analyze different subsets of replications that vary on these characteristics.

Principled statistical approach to evaluating replication evidence. The next step in our framework is to use a principled approach to statistically evaluate replication evidence. When only one or a few replications are available, replication evidence is evaluated at the individual-study level. When several replications are available, replication evidence is evaluated at both the individual-study and the meta-analytic levels. In this framework, a meta-analysis synthesizes evidence across replication studies nested within distinct generalizations of an effect (the original study's ES is not included in the meta-analytic estimate). Whether replication evidence is statistically evaluated at the individual-study level or meta-analytically, we propose a statistical approach that is more nuanced than what is currently standard practice in the field and that also uses clearer language to describe replication results (see also LeBel, Vanpaemel, Cheung, & Campbell, 2018). Three distinct statistical aspects of results are considered: (a) whether a signal was detected in the replication, (b) the consistency of the replication ES with the original study's ES, and (c) the precision of the replication's ES estimate. Consider, for example, the replication scenarios in Figure 2. In the case of Replication #1, one would say that a signal was detected (i.e., the confidence interval for the replication ES excludes zero) and that the replication ES is consistent with the original study's (i.e., the replication's confidence interval includes the original ES point estimate). This is the most favorable outcome of a severe falsifying test. In contrast, in Replications #2, #3, and #4, a signal was detected, but the replication ES is inconsistent with the original ES point estimate, a less favorable replication outcome suggesting that boundary conditions of the target effect may not yet be well understood. Finally, in Replications #5 and #6, the replication evidence is even less favorable given that no signal was detected, and Replication #6 represents the least favorable outcome: absence of a signal in combination with a replication ES estimate that is inconsistent with the original ES point estimate. When a replication ES estimate is less precise than the effect in the original study (i.e., the confidence interval for the replication ES is wider than the confidence interval in the original study), the label "imprecise" is added to warn readers that the replication result should only be interpreted meta-analytically.

In summary, an effect can be considered replicable when replications consistently detect a signal consistent with (i.e., of similar magnitude to) the ES point estimate from the original study (Replication #1). When several replication studies are available for a specific operationalization of an

original effect and a meta-analysis is conducted, an effect is considered replicable when the meta-analytic ES estimate excludes zero and is consistent with the original ES point estimate.

The Curate Science Web Platform

This proposed unified curation framework is currently guiding the design and implementation of a crowdsourced searchable Web platform, curatescience.org, that will allow the community of researchers to curate and evaluate the transparency, reproducibility, robustness, and replicability of each other's findings in an incremental, ongoing basis. A nonstatic Web platform is crucial because scientific evidence is dynamic and constantly evolving: New evidence can always count against, or be consistent with, a previously accepted hypothesis. In the digital era, it no longer makes sense to continue publishing literature reviews of evidence as static documents that become out-of-date shortly after they are submitted to a journal for peer review (as happens with traditional meta-analyses). This crowdsourced, incremental platform is decentralized, and thus the contributed evidence can (a) be inclusive, (b) originate from researchers with maximally diverse intellectual and theoretical viewpoints, and (c) be up-to-date.

The platform will allow users to search for (and filter) studies on the basis of characteristics related to transparency, reproducibility, robustness, and replicability. For example, researchers will be able to search for articles that (a) comply with minimum levels of different kinds of transparency (e.g., they may want to find only articles that report preregistered studies with open materials or only articles with publicly available data and reproducible code files), (b) report reproducibility or robustness reanalyses of published findings, or (c) report replications of published effects.

The platform will have several features for curating transparency. Researchers will be able to indicate that their studies already complied with a specific reporting standard (e.g., the basic-4 reporting standard) at the time of publication or to retroactively disclose unreported information so that their studies comply with a chosen standard. A standardized labeling system will be used to indicate whether a study complies with a reporting standard and, if so, which one. This feature is crucial given that only a minority of journals require compliance to such standards and those that do do not use a standardized labeling system.[6] Researchers will also be able to earn open-practice badges for studies published in journals that do not yet award these badges; the relevant badge icons will be hyperlinked to the URLs of the publicly available resources (i.e., open materials, preregistered protocols, open data, and reproducible code files; see Fig. 3).

The platform also will support the curation of reproducibility and robustness. Users will be able to add articles reporting reproducibility or robustness reanalyses (see Fig. 3). They will also be able to upload (and get credit for) verifications of the analytic reproducibility and robustness of a study's primary substantive finding. From the perspective of falsifiability, it is crucial that such verifications are themselves easily scrutinizable so that they can be verified by independent researchers (see Appendix D, Fig. 2, at https://osf.io/gpu3a/ for a screenshot showing how such verifications will be displayed in search results).

Finally, the platform also will support the curation of replicability. It will allow users to add articles reporting replications of published effects (see Fig. 3). It will also allow them to add replications to preexisting collections of replication evidence and to create new evidence collections for effects not yet available in the database. Within their own Web browsers, researchers will be able to meta-analyze the evidence provided by replications that they have selected on the basis of key curated study characteristics (e.g., methodological similarity, design differences, preregistration status; see Appendix D, Fig. 3, at https://osf.io/gpu3a/ for a screenshot showing how this information will be displayed).

The success of the platform will hinge on researchers' active involvement with the Web site and contributions to its content (e.g., adding missing replications, curating study information, performing reproducibility analyses). To incentivize contributions, and also to maximize the quality of the contributed content, we will include key features guided by principles of social accountability and reward.[7] For example, all of a user's contributions will be prominently displayed on his or her public profile page, and recent contributions will be conspicuously displayed on the home page (and will include the contributors' names, which can be clicked on to see those researchers' profile pages). To maximize the number and frequency of contributions, we will follow a "low barrier to entry," incremental approach, leaving as many fields optional as possible, so that the curation of information can be continued later by other users and editors. To maximize the quality of the contributed content, the platform will track the user name and date for all added and updated information and will also feature light-touch editorial review for certain categories of information (e.g., when a new replication study is added to an existing evidence collection, the information will be marked as "unverified" until another user or editor reviews it).

Example: The Infidelity-Distress Effect

To demonstrate our proposed framework, in this section we apply it to the original and replication studies of the infidelity-distress effect (Buss et al., 1999, Study 2; see

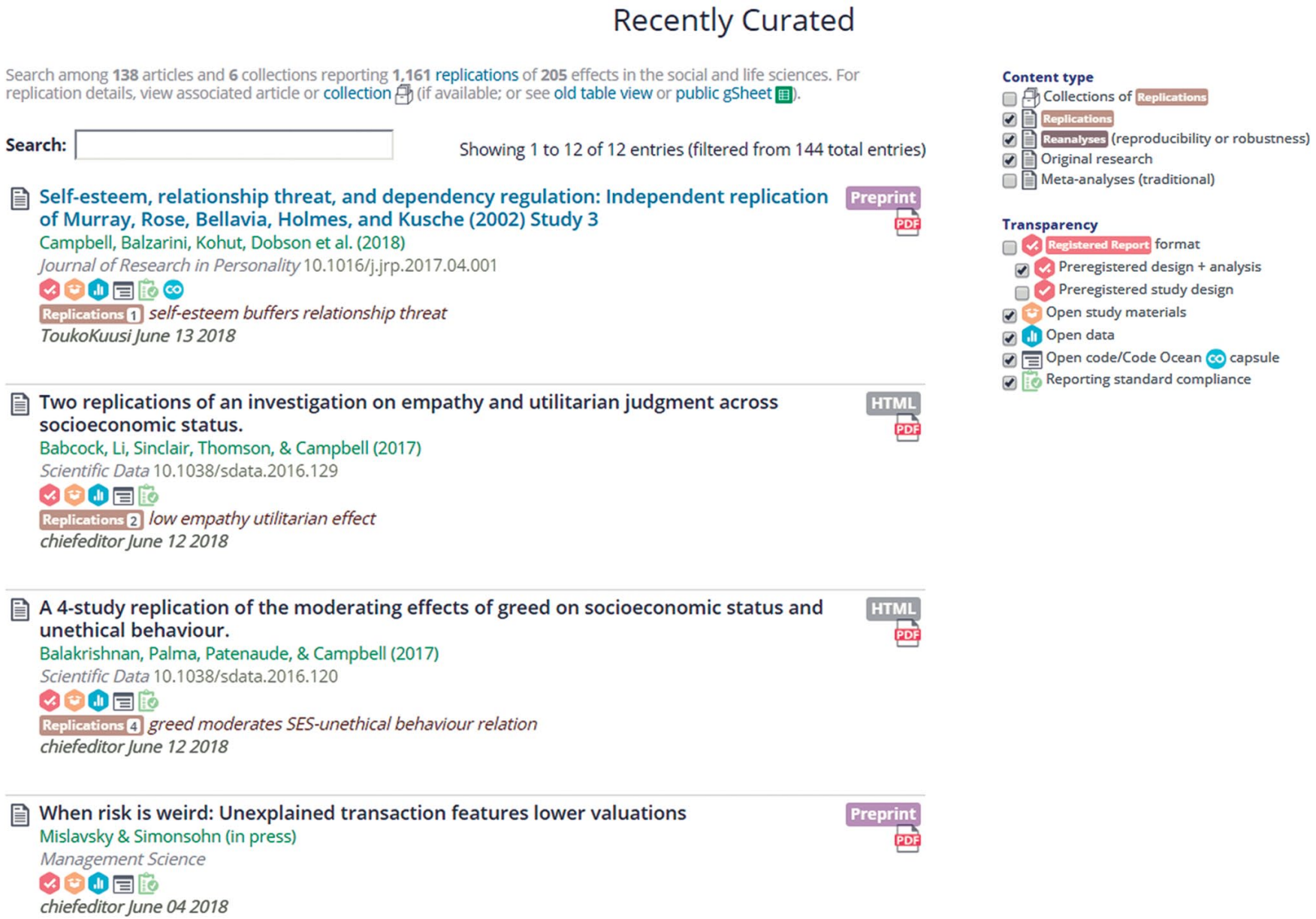

Fig. 3. Screenshot showing the list of recently curated articles on the Web platform's main page on July 18, 2018 (see https://curatescience.org). Articles can be filtered on the basis of their transparency, as indicated by the badge icons, which denote preregistration types; availability of study materials, data, and code; and compliance with reporting standards. Users can click on the icons to access an article's publicly available content. Articles can also be filtered by whether they report replications (note that the number of replications and the target effect are indicated) or reproducibility or robustness reanalyses.

Appendix E at https://osf.io/gpu3a/ for two additional examples). In the original study (Buss et al., 1999, Study 2) undergraduate students indicated whether they would be more distressed by a (hypothetical) sexual or emotional infidelity committed by their partner (forced choice). On average, men were more likely than women to report that the sexual infidelity would be more distressing. In a population generalization, the effect was generalized to an older community sample of individuals (i.e., mean age = 67.1 years; Shackelford et al., 2004).

Transparency

The reported methodological details of Buss et al.'s (1999) and Shackelford et al.'s (2004) studies did not meet the basic-4 or more comprehensive reporting standards, though the articles did comply with the reporting standards at the time. The studies do not qualify for open-practice badges, nor were they preregistered given that the research was conducted before the advent of such practice.

Analytic reproducibility and robustness

Because the data for the studies are not publicly available, verifications of their analytic reproducibility and robustness are not possible.

Effect replicability

We evaluated the replicability of the infidelity-distress effect on the basis of known eligible replication studies. Each included replication complied with the basic-4 reporting standard, had open materials and open data, and was also preregistered. The open data allow independent verifications of the analytic reproducibility of the reported results. Indeed, the first author of the present article attempted such a verification and was able to successfully reproduce the reported primary-outcome effect sizes (within a 10% margin of error) for all five of IJzerman et al.'s (2014) replications. The fact that the replications were preregistered helps rule out the possibility that more minor forms of analytic and design flexibility biased the results (assuming that the preregistration was sufficiently detailed and that the study procedures reported followed the preregistered protocol). Though no evidence of positive controls was reported, the open data make it possible to evaluate the plausibility of auxiliary hypotheses, by examining estimates of the internal consistency of individual differences that were assessed. For example, the measure of sociosexual orientation exhibited high internal consistency ($\alpha = .87$, $\alpha = .85$, $\alpha = .80$, and $\alpha = .86$ across IJzerman et al.'s Studies 1 through 4, respectively), which suggests that it is plausible that auxiliary hypotheses were sound. The design of these replication studies differed in several ways from the original studies: They were conducted in Dutch instead of English, Study 4 was conducted online instead of in the lab, and the infidelity-distress measure consisted of eight dilemmas (provided by the original authors to the researchers conducting the replications) instead of six. All the replications involved independent investigators.

As Figure 4 shows, for the original effect observed among young individuals, the meta-analytic replication evidence reveals an infidelity-distress effect, d, of 0.57, 95% confidence interval = [0.18, 0.96] (not including the original study's ES estimate). Thus, a signal was detected. However, the confidence interval for the meta-analytic ES estimate excludes the original study's ES point estimate of 1.30; hence, the meta-analytic result is considered inconsistent with the original study (as in Replication #3 in Fig. 2). This result suggests that the original study may have overestimated the effect's magnitude, that boundary conditions for the effect are still not well understood, or both. For the population generalization, no signal was detected; in addition, the meta-analytic ES estimate, d, of −0.01, 95% confidence interval = [–0.22, 0.20], is inconsistent with the original study's ES point estimate of 0.57 (see Fig. 4; also, cf. Replication #6 in Fig. 2). This result suggests that the effect may not generalize to older individuals.

Conclusion

We have proposed a unified framework for systematically quantifying the method and data transparency, analytic reproducibility, analytic robustness, and effect replicability of published scientific findings. The framework is unique among extant approaches in several ways. It is the only framework that integrates deep-level curation of transparency, reproducibility, robustness, and replicability of empirical research in a harmonized, flexible system that is logically ordered to maximize research efficiency. Specifically, it is unique in curating, at the study level, the transparency of published findings (i.e., compliance to reporting standards, public availability of materials and data, preregistration information) and in including standardized workflows and scoring procedures for estimating the degree of reproducibility and robustness of reported results. The framework also provides a novel system for organizing and evaluating the replicability of effects by curating key characteristics of replication studies so that replication results can be statistically evaluated in a nuanced manner at the meta-analytic and individual-study levels.

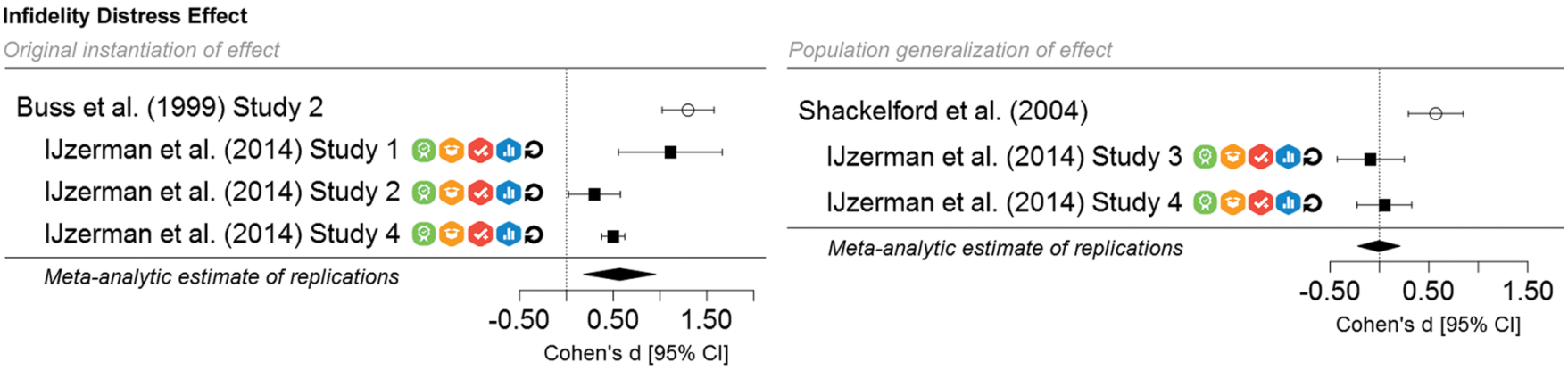

Fig. 4. Meta-analytic results for the replication studies investigating the original infidelity-distress effect (Buss et al., 1999, Study 2) and its population generalization (Shackelford et al., 2004). From left to right, the icons indicate that each replication study complied with a reporting standard, has open materials, was preregistered, has open data, and was confirmed to be analytically reproducible. For each study, the plots show the observed effect size (*d*) and its 95% confidence interval (CI).

Table 1. Benefits of Curating the Transparency, Reproducibility, Robustness, and Replicability of Empirical Research

Benefit category	Benefit
Theory building and application	• Researchers can base beliefs about the credibility of effects on empirical evidence rather than authority (e.g., journal or university prestige). • Researchers can more accurately estimate effect sizes within a research area, and thereby better estimate sample sizes needed to achieve sufficient statistical power. • Researchers can identify important studies that have not yet been replicated and commission such replications (via, e.g., StudySwap or the Psychological Science Accelerator).
Metascience	• Curation can yield a rich database of transparently reported original and replication studies that can be used for metascience research to deepen understanding of the predictors of replicability (e.g., the original study's *p* value, sample size, study design). • The curated information can be used to track transparency, reproducibility, robustness, and replicability of studies over time in order to gauge a discipline's progress in achieving higher research integrity.
Pedagogy	• A searchable database can be used to teach students about transparency and replication (e.g., it provides real-world examples of effects exhibiting different levels of replicability; it can also inform teachers about replicable effects that can justifiably be included in course materials).
Practical benefits	• Curated information can help researchers locate publicly available experimental materials for follow-up research and publicly available data sets and reproducible code for secondary analyses and reanalyses from alternative theoretical perspectives. • Researchers can identify replicable effects that are ready to be extended (which is particularly useful for graduate students, early-career researchers, and applied researchers).
Social norms	• Making it easier to find transparently reported research increases the likelihood that ambivalent or unaware researchers will decide to adopt transparent practices, and hence can accelerate a cultural shift in the research community so that it becomes the social norm to report one's research transparently. • Increasing the visibility of replication studies rewards researchers who devote their time to replicating the work of others.

In conclusion, it is important to mention what the unified framework, and its Web implementation, is *not* intended to be. It is not intended to provide a debunking platform aimed at cherry-picking unfavorable evidence regarding the replicability of published findings. It is also not intended to be a "final authoritative arbiter" of research quality. In contrast, it is a system for organizing scientific information and developing metascientific tools to help the community of researchers carefully evaluate research in a nuanced manner. It is also not a private club, but rather is an open, decentralized, and transparently accountable public resource available to all researchers who abide by the relevant scientific codes of conduct and norms of civil communication.

Crowdsourcing the credibility of published research creates value and is expected to lead to several distinct benefits, summarized in Table 1.

We hope that this article will serve as a call to action for the research community in psychology (and related disciplines) to get involved in using, designing, and contributing to the Web platform curatescience.org. The vision is that of a vibrant community of individuals who use and contribute to the platform in a collective bid to digitally organize the published literature. This crowdsourcing of the credibility of empirical research will accelerate theoretical understanding of the world as well as the development of applied solutions to society's most pressing social and medical problems.

Action Editor

Simine Vazire served as action editor for this article.

Author Contributions

E. P. LeBel conceived the general idea of this article, drafted and revised the manuscript, created the figures, and executed the analytic-reproducibility checks and meta-analyses for the application of the framework to the infidelity-distress effect. W. Vanpaemel provided substantial contributions to the conceptual development of the ideas presented. W. Vanpaemel, R. J. McCarthy, B. D. Earp, and M. Elson provided critical commentary and made substantial contributions to writing and revising the manuscript. All the authors approved the final submitted version of the manuscript.

ORCID iDs

Etienne P. LeBel https://orcid.org/0000-0001-7377-008X
Wolf Vanpaemel https://orcid.org/0000-0002-5855-3885

Acknowledgments

We would like to thank E.-J. Wagenmakers, Rogier Kievit, Rolf Zwaan, Alexander Aarts, and Touko Kuusi for valuable feedback on earlier versions of this manuscript.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

M. Elson is supported by the Digital Society research program funded by the Ministry of Culture and Science of North Rhine-Westphalia, Germany.

Notes

1. All else being equal, a finding reported with lower levels of transparency should be considered less credible than a finding reported with greater transparency even if the lack of transparency is due to ethical constraints (e.g., participants' privacy, confidentiality issues). However, such a finding could nonetheless be considered credible if independent researchers can consistently replicate it in new samples.
2. Exceptions may sometimes apply, depending on the nature of the study. For example, although assessing replicability is normally the last step, for inexpensive and easy-to-implement cognitive-psychology studies, it may make sense to evaluate replicability without first gauging analytic reproducibility (though even in this scenario, a study's methodological details should first be thoroughly scrutinized, which requires sufficient method transparency).
3. One should not, however, conflate mere compliance with a reporting standard with high levels of methodological rigor.
4. Given that within our framework, studies need to be sufficiently methodological similar to an original study in order to be considered replication studies, they can be construed as tacitly "preregistered." However, formally preregistering design and analytic plans of replication studies can nonetheless further constrain more minor forms of design and analytic flexibility.
5. Unless preceded by a modifier (e.g., *far*), we use the term *replication* to refer to direct replications and *generalization* to refer to conceptual replications.
6. The platform will also eventually allow researchers to leave comments regarding methodological issues identified for a study (and will also allow them to add hyperlinks to other published commentaries and critiques about the study, e.g., from pubpeer.com or blog posts).
7. To further encourage contributions, and as is standard for crowdsourced platforms, during initial phases, we will pay (Ph.D.-level) curators to contribute content that will seed the database to sufficient levels to convince other users that the platform is wide-ranging enough to be worth contributing to. As of July 2018, the Web site features 1,161 partially curated replication studies on 205 effects from the cognitive- and social-psychology literatures.

References

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., . . . Stroup, D. F. (1996). Improving the quality of reporting of randomized controlled trials: The CONSORT statement. *Journal of the American Medical Association*, *276*, 637–639.

Buss, D. M., Shackelford, T. K., Kirkpatrick, L. A., Choe, J. C., Lim, H. K., Hasegawa, M., . . . Bennett, K. (1999). Jealousy and the nature of beliefs about infidelity: Tests of competing hypotheses about sex differences in the United States, Korea, and Japan. *Personal Relationships*, *6*, 125–150.

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

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis*. New York, NY: Russell Sage Foundation.

Doyen, S., Klein, O., Simons, D. J., & Cleeremans, A. (2014). On the other side of the mirror: Priming in cognitive and social psychology. *Social Cognition*, *32*, 12–32.

Earp, B. D. (in press). Falsification: How does it relate to reproducibility? In J.-F. Morin, C. Olsson, & E. O. Atikcan (Eds.), *Key concepts in research methods*. Abingdon, England: Routledge.

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, Article 621. doi:10.3389/fpsyg.2015.00621

Eich, E. (2014). Business not as usual. *Psychological Science*, *25*, 3–6.

Feynman, R. P. (1974). Cargo cult science. *Engineering and Science*, *37*(7), 10–13.

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., . . . Frank, M. C. (2018). *Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition*. Retrieved from https://osf.io/preprints/bitss/39cfb/

Hendrick, C. (1991). Replication, strict replications, and conceptual replications: Are they important? In J. W. Neuliep (Ed.), *Replication research in the social sciences* (pp. 41–49). Newbury Park, CA: Sage.

Higgins, J. P. T., Lasserson, T., Chandler, J., Tovey, D., & Churchill, R. (2018). *Methodological Expectations of Cochrane Intervention Reviews (MECIR): Standards for the conduct and reporting of new Cochrane Intervention Reviews, reporting of protocols and the planning, conduct and reporting of updates*. Retrieved from http://community.cochrane.org/sites/default/files/uploads/MECIR%20PRINTED%20BOOKLET%20FINAL%20v1.02.pdf

IJzerman, H., Blanken, I., Brandt, M. J., Oerlemans, J. M., Van den Hoogenhof, M. M. W., Franken, S. J. M., & Oerlemans, M. W. G. (2014). Sex differences in distress from infidelity in early adulthood and in later life: A replication and meta-analysis of Shackelford et al. (2004). *Social Psychology*, *45*, 202–208.

Kashy, D. A., Donnellan, M. B., Ackerman, R. A., & Russell, D. W. (2009). Reporting and interpreting research in PSPB: Practices, principles, and pragmatics. *Personality and Social Psychology Bulletin*, *35*, 1131–1142.

Kidwell, M. C., Lazarevic´, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., . . . Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLOS Biology*, *14*(5), Article e1002456. doi:10.1371/journal.pbio.1002456

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91–196). Cambridge, England: Cambridge University Press.

LeBel, E. P., Berker, D., Campbell, L., & Loving, T. J. (2017). Falsifiability is not optional. *Journal of Personality and Social Psychology*, *113*, 254–261.

LeBel, E. P., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K. R., Ratliff, K. A., & Smith, C. T. (2013). PsychDisclosure.org: Grassroots support for reforming reporting standards in psychology. *Perspectives on Psychological Science*, *8*, 424–432.

LeBel, E. P., & John, L. (2017). Psychological and institutional obstacles toward more transparent reporting of psychological science. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. 73–84). New York, NY: John Wiley & Sons.

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology*, *15*, 371–379.

LeBel, E. P., Vanpaemel, W., Cheung, I., & Campbell, L. (2018). *Evaluating replications with nuance*. Retrieved from https://osf.io/paxyn/

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103–115. doi:10.1086/288135

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834. doi:10.1037/0022-006X.46.4.806

Moery, E., & Calin-Jageman, R. J. (2016). Direct and conceptual replications of Eskine (2013): Organic food exposure has little to no effect on moral judgments and prosocial behavior. *Social Psychological & Personality Science*, *7*, 312–319.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, USA*, *115*, 2600–2606.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*, 531–536.

Popper, K. R. (1959). *The logic of scientific discovery*. London, England: Hutchinson.

Rosenthal, R. (1991). Replication in behavioral research. In J. W. Neuliep (Ed.), *Replication research in the social sciences* (pp. 1–39). Newbury Park, CA: Sage.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*, 90–100.

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *BMC Medicine*, *8*, Article 18. doi:10.1186/1741-7015-8-18

Shackelford, T. K., Voracek, M., Schmitt, D. P., Buss, D. M., Weekes-Shackelford, V. A., & Michalski, R. L. (2004). Romantic jealousy in early adulthood and in later life. *Human Nature*, *15*, 283–300.

Simmons, J., Nelson, L., & Simonsohn, U. (2012). A 21 word solution. *Dialogue: The Official Newsletter of the Society for Personality and Social Psychology*, *26*(2), 4–7.

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.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). *Specification curve: Descriptive and inferential statistics on all reasonable specifications*. Retrieved from http://ssrn.com/abstract=2694998

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*, 702–712.

Vandenbroucke, J. P., von Elm, E., Altman, D. G., Gøtzsche, P. C., Mulrow, C. D., & Pocock, S. J., . . . STROBE Initiative. (2014). Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and elaboration. *International Journal of Surgery*, *12*, 1500–1524.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2017). Making replication mainstream. *Behavioral & Brain Sciences*. Advance online publication. doi:10.1017/S0140525X17001972