

Tracking replicability as a method of post-publication open evaluation

Joshua K. Hartshorne*, Adena Schachner

Department of Psychology, Harvard University, Cambridge, MA, USA

Send Correspondence to:

Joshua Hartshorne
33 Kirkland Street
Cambridge, MA 02138 USA
jharts@wjh.harvard.edu

Running title: Tracking replicability

Word count: 7,507 words + Appendix (695 words)

Acknowledgements:

The first author was supported through the National Defense Science and Engineering Graduate Fellowship (NDSEG) Program. Many thanks to Tim O'Donnell, Manizeh Khan, Tim Brady, Roman Feiman, Jesse Snedeker and Susan Carey for discussion and feedback.

Abstract

Recent reports have suggested that many published results are unreliable. To increase the reliability and accuracy of published papers, multiple changes have been proposed, such as changes in statistical methods. We support such reforms. However, we believe that the incentive structure of scientific publishing must change for such reforms to be successful. Under the current system, the quality of individual scientists is judged on the basis of their number of publications and citations, with journals similarly judged via numbers of citations. Neither of these measures takes into account the replicability of the published findings, as false or controversial results are often particularly widely cited. We propose tracking replications as a means of post-publication evaluation, both to help researchers identify reliable findings and to incentivize the publication of reliable results.

Tracking replications requires a database linking published studies that replicate one another. As any such database is limited by the number of replication attempts published, we propose establishing an open-access journal dedicated to publishing replication attempts. Data quality of both the database and the affiliated journal would be ensured through a combination of crowd-sourcing and peer review. As reports in the database are aggregated, ultimately it will be possible to calculate replicability scores, which may be used alongside citation counts to evaluate the quality of work published in individual journals. In this paper, we lay out a detailed description of how this system could be implemented, including mechanisms for compiling the information, ensuring data quality, and incentivizing the research community to participate.

Key words: replication, replicability, post-publication evaluation, open evaluation

Improving the Quality of Published Research

The current system of conducting, reviewing and publishing scientific findings -- while enormously successful -- is by no means perfect. Peer review, the primary vetting procedure for publication, is often slow, contentious and uneven (Cole Jr., Cole, & Simon, 1981; Eysenck & Eysenck, 1992; Mahoney, 1977; Newton, 2010; Peters & Ceci, 1982). Incorrect use of inferential statistics leads to publication of spurious findings (Baayen, Davidson, & Bates, 2008; Jaeger, 2008; Saxe, Brett, & Kanwisher, 2006; Vul, Harris, Winkielman, & Pashler, 2009; Wagenmakers, Wetzels, Borsboom, & Van, in press). Publication biases, such as the bias against publishing null results (e.g., Boffetta, et al., 2008; Easterbrook, Berlin, Gopalan, & Matthews, 1991; Ioannidis, 2005b), lead to distortions in the published record, hampering both informal reviews and formal meta-analyses. Numerous valuable proposals have been offered as to how to improve the system in order to enable researchers to better identify high-quality research, including those in the present special issue.

There are many considerations that go into determining research quality, but perhaps the most fundamental is replicability. Recently, numerous reports have suggested that many published results across a range of scientific disciplines do not replicate (Boffetta, et al., 2008; Ferguson & Kilburn, 2010; Ioannidis, 2005a; Ioannidis, Ntzani, Trikalinos, & Contopoulos-Ioannidis, 2001; Jennions & Møller, 2002b; Lohmueller, Pearce, Pike, Lander, & Hirschhorn, 2003). However, because replication attempts are not tracked and are often not reported, there is no systematic way for researchers to know which results in the literature have been replicated.

In the present paper, we first discuss evidence that the rate of replicability of published studies is low, including novel data from a survey of researchers in psychology and related fields. We propose that this low replicability stems from the current incentive structure, in which

replicability is not systematically considered in measuring paper, researcher and journal quality. As a result, the current incentive structure rewards the publication of non-replicable findings, complicating the adoption of needed reforms. Thus, we outline a proposal for tracking replications as a form of post-publication evaluation, and using these evaluations to calculate a metric of replicability. In doing so, we aim not only to enable researchers to easily find and identify reliable results, but also to improve the incentive structure of the current system of scientific publishing, leading to widespread improvements in scientific practice and increased replicability of published work.

Why might we expect low replicability?

Many aspects of current accepted practice in psychology, neuroscience and other fields necessarily decrease replicability. Some of the most common issues include a lack of documentation of null findings; a tendency to conduct low-powered studies; failure to account for multiple comparisons; data-peeking (with continuation of data collection contingent on current significance level); and a publication bias in favor of surprising ("newsworthy") results.

Lack of publication or documentation of null findings

Null results are less likely to be published than statistically significant findings. This has been extensively documented in the medical literature (Callaham, Wears, Weber, Barton, & Young, 1998; Dickersin, Chan, Chalmers, Sacks, & Smith, 1987; Dickersin, Min, & Meinert, 1992; Dwan, et al., 2008; Easterbrook, et al., 1991; Misakian & Bero, 1998; Olson, et al., 2002; Sena, Worp, Bath, Howells, & Macleod, 2010), with additional reports in political science (Gerberg, Green, & Nickerson, 2001), ecology and evolution (Jennions & Møller, 2002a), and clinical psychology (Coursol & Wagner, 1986; Cuijpers, Smit, Bohlmeijer, Hollon, & Andersson, 2010). There appear to be fewer comprehensive studies of publication bias in non-

clinical psychology, although evidence of this bias has been documented in a few specific literatures (Ferguson & Kilburn, 2010; Field, Munafo, & Franken, 2009).

Preferential publication of significant effects necessarily biases the record. Consider cases in which multiple labs all test the same question, or in which the same lab repeatedly tests the same question while iteratively refining the method. By chance alone, some of the experiments will result in publishable statistically significant effects; the likelihood that a finding may be spurious is masked by the fact that the null results are not published.

The significance-bias also leads to the overestimation of real effects. Measurement is probabilistic: the measured effect size in a given experiment is a function of the true effect size plus some random error. In some experiments, the measured effect will be larger than the true effect, and in some it will be smaller. Suppose the statistical power of the experiment is 0.8 (a particularly high level of power for studies in psychology; see below). This means that the effect will be statistically significant only if it is in the top 80% of its sampling distribution. 20% of the time, when the effect is -- by chance -- relatively small, the results will be non-significant. Thus, given that an effect was significant, the measured effect size is probably larger than the actual effect size, and subsequent measurements will find smaller effects due to the familiar phenomenon of regression to the mean. The lower the statistical power, the more the effect size will be inflated.

Low power, small effect size

A number of findings suggest that the statistical power in psychology and neuroscience experiments is typically low. According to multiple meta-analyses, the statistical power of a typical psychology or neuroscience study to detect a medium-sized effect (defined variously as $r=.3$, $r=.4$ or $d=.5$) is approximately .5 or below (Bezeau & Graves, 2001; Cohen, 1962;

Kosciulek & Szymanski, 1993; Sedlmeier & Gigerenzer, 1989). In applied psychology, power for medium effects is closer to .7, though it remains low for small effects (Chase & Chase, 1976; Mone, Mueller, & Mauland, 1996; Shen, et al., in press). Nonetheless, many effects of interest in psychology are small and thus typical statistical power may be quite low. Field, Munafò and Franken (2009) report an average power of .2 in a meta-analysis of 68 studies of craving in addicts and attentional bias. In a heroic meta-analysis of 322 meta-analyses in social psychology, Richard, Bond and Stokes-Zoota (2003) report that the average effect size was $r=.21$. To achieve power of .8 would require the average study to have 173 participants (in terms of medians: $r=.18$, $N=237$), already far larger than typical sample size. Nearly 1/3 of the effect sizes reported were $r=.1$ or less, requiring $N=772$ to achieve power of .8.

All else being equal, low statistical power would increase the proportion of significant results that are spurious. For instance, suppose researchers are investigating a hypothesis that is equally likely to be true or false (the prior likelihood of the null hypothesis is 50%), using methods with statistical power=0.8. In this case, 6% of significant results will be false positives (True positives: $0.5*0.8=0.4$; False positives: $0.5*0.05=0.025$; Ratio: $0.025/0.425=0.059$). If Power=0.2, this increases to 20%. If the prior likelihood of the null hypothesis is 90% (i.e., if an effect would be surprising, or when data-mining), the false positive rate will be 69% (for additional discussion, see Yarkoni, 2009; for other problems associated with small power, see Tversky & Kahneman, 1971).

Failure to account for multiple comparisons

If one tests for 10 different possible effects in each experiment, the chance of finding at least one significant at the $p=.05$ level even when no effect actually exists is $1 - 0.95^{10} = 0.4$. Since experiments with large numbers of comparisons are often entirely exploratory, where there

is no strong *a priori* reason to believe that any of the investigated effects exist, the false positive rate may approach 100% for data-mining studies with large datasets.

Data-peeking and contingent stopping of data collection

Many researchers compile and analyze data prior to testing a full complement of subjects. There is nothing wrong with this, so long as the decision to stop data collection is made independent of the results of these preliminary analyses, or so long as the final result is then replicated with the same number of subjects. Unfortunately, the temptation to stop running participants once significance is reached -- or to run additional participants if it has not been reached -- is difficult to resist. This data peeking and contingent stopping has the potential to significantly increase the false positive rate (Armitage, McPherson, & Rowe, 1969; Feller, 1940; Yarkoni & Braver, 2010). A researcher who tests for significance after every participant has a 25% chance of finding a significant result with 20 or fewer participants (if the underlying distribution is normal; the analogous numbers are 19.5% for exponential distributions and 11% for binomial distributions; Armitage et al., 1969). This issue may be mitigated by use of alternative statistical tests, such as Bayesian statistics (Edwards, Lindman, & Savage, 1963), but such statistics have not been widely adopted.

Newsworthiness bias

Researchers are more likely to submit -- and editors more likely to accept -- "newsworthy" or surprising results. Spurious results are likely to be surprising, and thus are likely to be over-represented in published reports. Consistent with this claim, there is some evidence that highly-cited papers are less likely to replicate (Ioannidis, 2005a) and that publication bias affects high-impact journals more severely (Ioannidis, 2005a; Munafò, Stothart, & Flint, 2009).

How Replicable Are Published Studies?

Several studies have found low rates of replicability across multiple scientific fields. Ioannidis (2005) found that of 34 highly-cited clinical research studies for which replication attempts had been published, seven (20%) did not replicate. Boffetta et al. (2008) report a number of cases in which reports of significant cancer risk factors did not replicate. Recent studies have reported that relatively few genetic association links can be replicated (Hirschhorn, Lohmueller, Byrne, & Hirschhorn, 2002; Ioannidis, et al., 2001; Ioannidis, Trikalinos, Ntzani, & Contopoulos-Ioannidis, 2003; Lohmueller, et al., 2003; Trikalinos, Ntzani, Contopoulos-Ioannidis, & Ioannidis, 2004).

Likewise, several studies have found that initial reports of effect size are often exaggerated. This has been noted in medicine (Ioannidis, 2005a; Ioannidis, et al., 2001; Ioannidis, et al., 2003; Trikalinos, et al., 2004; but see Gehr, Weiss & Porzolt, 2006), with similar declines in effect size reported in ecological and evolutionary biology (Jennions & Møller, 2002a, 2002b). In the most extreme example, Dwald, Thursby and Anderson (1986) reanalyzed the datasets underlying published studies in economics and were unable to fully replicate the analyses for seven of nine (78%).

Less is known about replication rates in psychology and neuroscience. In a series of five meta-analyses of fMRI studies, Wager and colleagues estimated that between 10% and 40% of activation peaks are false-positives (Wager, Lindquist, & Kaplan, 2007; Wager, Lindquist, Nichols, Kober, & van Snellenberg, 2009). While there seem to be few systematic surveys within psychology, some published effects are known not to replicate, such as the initial finding that violent video games increase violent behavior (Ferguson & Kilburn, 2010), various claims about the relationship between birth order and personality (Ernst & Angst, 1983; Harris, 1998; but see:

Hartshorne, Salem-Hartshorne, & Hartshorne, 2009; Kristensen & Bjerkedal, 2007), and a range of gene/environment interactions (Flint & Munafò, 2009).

In order to add to our knowledge of replicability rates in psychology and related disciplines, we surveyed 49 researchers in these disciplines, who reported a total of 257 attempted replications of published studies (for details, see Appendix). Only 127 (49%) fully replicated the original findings. This low rate was not driven by a small number of researchers attempting a large number of poor quality replications: both the mean and median replication success rates were 50%, with 77% of researchers reporting at least one attempted replication. Thus, the results of this survey suggest that replication rates within psychology and related disciplines are undesirably low, in accordance with the low rates of replicability found in many other fields.

Incentives in Publication

As reviewed above, a number of factors promote low replicability rates across a range of fields. These problems are reasonably well known, and in many cases solutions have been proposed, such as use of different statistical methods and self-replication prior to publication. However, in spite of these solutions, evidence suggests that replicability remains low and thus that the proposed solutions have not been widely adopted. Why would this be the case? We propose that the incentive structure of the current system diminishes the ability and tendency of researchers to adopt these solutions. Namely, current methods of judging paper, researcher and journal quality fail to take replicability into account, and in effect incentivize publishing spurious results.

Quantifying research quality

There are three primary *quantitative* criteria by which researchers are judged: their number of publications, the impact factor of the journals in which the publications appear, and the number of citations those papers receive. These quantitative values are a major consideration in the awarding of grants, hiring and tenure. Journals are similarly judged in terms of citation counts, which are compiled to calculate journal impact factors. Unfortunately, these metrics of quality tend to disincentivize taking additional steps to ensure the reliability of published findings, for several reasons.

Firstly, eliminating false positives means publishing fewer papers, since null results are difficult to publish. Second, ensuring that effect sizes are not inflated means reporting results with smaller effect sizes, which may be seen as less interesting or less believable. Third, as discussed above, spurious results are more likely to be surprising and newsworthy. Thus, eliminating spurious results disproportionately eliminates publications that would be widely-cited and published in top journals.

These drawbacks are compounded by the fact that many of the improved practices that ensure replicability take time and resources. Learning to use new statistical methods often requires substantial effort. Increasing an experiment's statistical power may require testing more participants. Eliminating stopping of data collection contingent on significance level (data-peeking) also means erring on the side of testing more participants. Perhaps the best insurance against false positives is pre-publication replication by the authors. All these strategies take time.

In addition, there is relatively little cost associated with publishing unreliable results, as failures to replicate are rarely published and not systematically tracked. As a result, knowledge of the replicability of results mainly travels via word-of-mouth, through specific personal interactions at conferences and meetings. There are obvious concerns about the reliability of such

a system, and there is little evidence that this system is particularly effective. We are aware of several cases in which a researcher invested months or years into unsuccessfully following up on a well-publicized effect from a neighboring subfield, only to later be told that it is "well-known" that the effect does not replicate.

Moreover, even when a failure-to-replicate is published, the results often go unnoticed. For example, a meta-analysis by Maraganore et al. (2004) concluded that UCHL1 is a risk-factor for Parkinson's Disease. Subsequent more highly-powered meta-analyses overturned this result (Healy et al., 2006). Nonetheless, Maraganore et al. (2004) has been cited 70 times since 2007 (Google Scholar, 5/10/2011), much to the dismay of the senior author of the study (Ioannidis, 2011). Even papers retracted by the authors remain in circulation. In 2001, [two papers were retracted](#) by Karen Ruggiero (Ruggiero & Marx, 1999; Ruggiero, Steele, Hwang & Marx, 2000). Nonetheless, 10 of the 22 citations to these papers were made in 2003 or later (Google Scholar, 4/25/2011). Similarly, though Lerner [requested the retraction](#) of Gonzalez & Lerner (2005) in 2008, the paper has been cited 5 times in 2010-2011 (Google Scholar, 4/25/2011).

It follows that researchers who take additional steps to ensure the quality of their data will ultimately spend more time and resources on each publication and, all else equal, will end up with fewer, less-often-cited papers in lower-quality journals. In the same way, journals that adopt more stringent publication standards may drive away submissions, particularly of the surprising, newsworthy findings that are likely to be widely-cited. Certainly, the vast majority of researchers and editors are internally motivated to publish real, reliable results. However, we also cannot continue practicing science without jobs, grants and tenure. This situation sets up a classic Tragedy of the Commons (Hardin, 1968): While it is in everyone's collective interest to adopt

strategies to improve replicability, the incentives for any *individual* researcher run the other direction.

Escaping the Tragedy of the Commons

Individuals can solve the Tragedy of the Commons by adopting common rules or changing incentive structures. To give a recent example, Jaeger (2008), Baayen (2008) and others convinced many language processing researchers to switch from ANOVAs to mixed effects models, in part by convincing editors and reviewers to insist on it. In this case, collective action motivated widespread adoption of an improved method of analysis.

In a similar way, collective action is needed to solve the problem of low replicability: Because the incentive structure of the current system penalizes any member of the community who is an early adopter of reforms, an organized community change is needed. Instead of maintaining a system in which individual incentives (publish as often as possible) run counter to the goals of the group (maintain the integrity of the scientific literature), we can change the incentives by placing value on replicability directly. To do this, we propose tracking the replicability of published studies, and evaluating the quality of work post-publication partly on this basis. By tracking replicability, we hope to provide concrete incentives for improvements in research practice, thus allowing the widespread adoption of these improved practices.

Replication Tracker: A Proposal

Below, we lay out a proposal for how replications might be tracked via an online open-access system tentatively named *Replication Tracker*. The proposed system is not yet constructed; our aim in this proposal is to spur necessary discussion on the implementation of such a system. We first describe the core components of such a system. We then discuss in more

depth issues that arise, such as motivating participation, aggregating information and ensuring data quality.

Core elements of the Replication Tracker

In a system such as Google Scholar, each paper's reference is presented alongside the number of times that paper has been cited, and each paper is linked to a list of the papers citing that target paper. Replication Tracker would function in a similar manner, except that it would be additionally indexed by specialized citations that link papers based on one attempting to replicate the other. Thus, each paper's reference would appear alongside not only a citation count, but an attempted replication count and information about the paper's replicability.

Replication Tracker's attempted replication citations are termed *Replication Links* (henceforth *RepLinks*). Each RepLink is tagged with metadata, answering the question: To what extent are these findings strong evidence that the target paper does or does not replicate? This metadata takes the form of two numerical ratings: a *Type of Finding Score*, running from +2 (fully replicated) to -2 (fully failed to replicate); and a *Strength of Evidence Score*, running from 1 (weak evidence) to 5 (strong evidence). These ratings, as well as the RepLinks themselves, could be produced through a variety of methods; we suggest crowd-sourcing from the scientific community, as outlined below.

For replications to be tracked, they must be reported. As discussed above, many replication attempts remain unpublished. Thus, Replication Tracker would be paired with an online, open-access journal devoted to publishing Brief Reports of replication attempts. After a streamlined peer review process, these Brief Reports would be published and connected to the papers they replicate via RepLinks in the Replication Tracker.

This system will ultimately form a rich dataset, consisting of RepLinks between attempted replications and the original findings. Each RepLink's ratings would indicate the type and strength of evidence of the findings. These ratings would be aggregated, and used to compute statistics on replicability. For instance, the system could summarize the data for each paper in terms of a *Replicability Score* (e.g., 15 attempted replications, Replicability Score: +1.7 (Partial Replication), Strength of Evidence: 4 (Strong)), much as citation indices score papers based on citation counts (e.g. cited by 15). These numbers would allow researchers to both get an initial impression of a finding's replicability at a glance, and quickly click through to the original sources for further detail. In addition, Replicability Scores could be aggregated for each journal, which could be used alongside the existing Impact Factor to evaluate the quality of journals.

Structure and content of RepLinks

RepLinks must, minimally, link a replication attempt with its target paper, note whether the finding was replication or non-replication, and note the strength of evidence for this finding.

There are many factors that enter into these decisions. For instance, a particular attempted replication may have investigated all of the findings in the target paper, or may have only attempted to replicate some subset. The findings may be more similar or less similar as well: All effects may have successfully replicated, or none; or some findings may have replicated while others did not. In addition, whether a replication serves as strong evidence of the replicability or non-replicability of the original finding depends on the extent of similarity of the methods used, and whether the attempt had high or low statistical power.

We propose capturing these issues in two ratings. The first rating, termed the *Type of Finding* rating, would take into account two factors: Whether all or only a subset of the target papers' findings were investigated; and whether all, none, or some of the attempted replications

were successful. On this Type of Finding scale, -2 would denote a total non-replication (all findings investigated; none replicated); -1 a partial non-replication (some subset of findings investigated; none of those investigated replicated); 0 would denote mixed results (of the findings investigated, some replicated and others did not); 1 a partial replication (some subset of findings investigated; all of those investigated replicated); and 2 a total replication (all findings investigated; all replicated).

The second rating would be a *Strength of Evidence* rating, scored on a 1 to 5 scale. This rating would take into account the remaining two factors: the extent to which the methods are similar between the target paper and the RepLinked paper, and the power of the replication attempt. Thus a score of 5 reflects a high-powered attempt with as-close-as-possible methods, while 1 reflects a low-powered attempt with relatively dissimilar methods. When a replication attempt is extremely low-power or uses substantially different methods, it would not be assigned a RepLink at all.

Who creates and rates RepLinks?

The ratings described above involve a number of difficult determinations. Given that no two studies can have exactly identical methods, how similar is similar enough? How does one determine whether a study has sufficient statistical power, given that the effect's size is itself under investigation?

To make these determinations, we turn to those individuals most qualified to make them: researchers in the field. Crowd-sourcing has proven a highly effective mechanism of making empirical determinations in a variety of domains (Bederson, Hu, & Resnik, 2010; Doan, Ramakrishnan, & Halevy, 2011; Franklin, Kossmann, Kraska, Ramesh, & Xin, 2011; Giles, 2005; Law, von Ahn, Dannenberg, & Crawford, 2007; von Ahn & Dabbish, 2008; von Ahn,

Maurer, McMillen, Abraham, & Blum, 2008; Yan, Kumar, & Ganesan, 2010). Researchers would form the user base of the system, and any user could submit a RepLink, as well as a Type of Finding and Strength of Evidence score for a RepLink. When submitting these materials, users could also optionally comment on each RepLink, providing a more detailed description of how the methods or results of the RepLinked paper differed from the target paper, or offering interpretations of discrepancies. These comments would be optionally displayed alongside each users' individual ratings, for readers looking for additional detail (Figure 4).

The system also utilizes multiple moderators. These moderators would take joint responsibility for tending the RepLinks and Brief Reports (see below) on papers in their sub-fields. Moderators would be scientists, and could be invited (e.g. by the founding members), although anyone with publications in the field could apply to be a moderator.

In submitting and rating RepLinks, researchers may disagree with one another as to the correct Type of Finding or Strength of Evidence ratings for a given RepLink, or may disagree as to whether two papers are sufficiently similar as to quality as a replication attempt. Users who agree with an existing rating may easily second it with a thumbs-up, while users who disagree with the existing ratings may submit their own additional ratings. Users who believe that the papers in question do not qualify as replications may flag the RepLink as irrelevant (RepLinks that have been flagged a sufficient number of times would no longer be used to calculate Replicability Scores, though these suppressed RepLinks would be visible under certain search options). These ratings would be combined together using crowd-sourcing techniques to determine the aggregate Type of Finding and Strength of Evidence scores for a given RepLink (see below).

Aggregation, Authority, and Machine Learning

Data must be aggregated by this system at multiple levels. First, multiple ratings for a given RepLink must be combined into aggregate Type of Finding and Strength of Evidence ratings for that RepLink. Second, where a single target paper has been the subject of multiple replication attempts, the different RepLinks must be aggregated into a single Replicability Score and Strength Score for that target paper. In the same way, scores may be combined across multiple papers to determine aggregate replicability across a literature, an individual researcher's publications, or a journal.

Aggregates need not be mere averages. How to best aggregate ratings across multiple raters is an active area of research in machine learning (Adamic, Zhang, Bakshy, & Ackerman, 2008; Albert & Dodd, 2004; Callison-Burch, 2009; Snow, O'Conner, Jurafsky, & Ng, 2008; Welinder, Branson, Belongie, & Perona, 2010). Type of Finding ratings for an individual RepLink may be weighted by their associated Strength of Evidence scores, as well as how many thumbs-up they have received.

In addition, ratings from certain users would be weighted based more heavily than others, as is done in many rating aggregation algorithms (e.g., Snow, et al., 2008). There are many mechanisms for doing so, such as downgrading the authority of users whose RepLinks are frequently flagged as irrelevant, and assigning greater authority to moderators. The best system of weighting and aggregating RepLinks is an interesting empirical question. We see no reason it must be set in stone from the outset; the best algorithms may be determined through new research in machine learning. To that end, the raw rating dataset would be made available to those working in machine learning and related fields.

A Note on Converging Results

Only strict replications, not convergent data from different methods, will be tracked in the proposed system. This may seem counter-intuitive, since tracking converging results is crucial for determining which theories are most predictive. However, the goal of the proposed system is not to directly evaluate which *theories* are right, but to determine which *results* are right—that is, which patterns of data are reliable. Consider that while converging results may suggest that the original finding replicates, *diverging* results may only indicate that the differences in the methodologies were meaningful. For this reason, we focus solely on tracking strict replications. We believe that evaluating the complex theoretical implications of a large body of data is best handled by researchers themselves (i.e. when writing review papers), and is likely not feasible with an automated system.

Authentication and labeling of authors' ratings and comments

Registering for the system and submitting RepLinks would not require authenticating one's identity. However, authors of papers could choose to have their identities authenticated in order to have comments on their own papers be marked as author commentaries (many RepLinks will almost certainly be submitted by authors, as they are most invested in the issues involved in replication of their own studies).

Identity authentication could be accomplished in multiple ways. For instance, a moderator could use the departmental website to verify the author's email address and send a unique link to that email address. Clicking on that link would enable the user to set up an authenticated account under the users' own name. Moderator's identities could be authenticated in a similar manner.

Retractions

The Replication Tracker system is also ideally suited to tracking retractions. Retractions may be submitted by users as a specially-marked type of RepLink, which would require moderator approval before posting. Retracted studies would appear with the tag *RETRACTED* in any search results, and automatically be excluded from calculations of Replicability Scores. As a safeguard against incorrect flags, any time a study is flagged as retracted, all other moderators would be notified, and the flag could be revoked if found inaccurate.

Brief Reports

The efficacy of Replication Tracker is limited by the number of published replication attempts. As discussed above, both successful replications and null results are difficult to publish, and often remain undocumented. Thus, we propose launching an open-access journal that publishes all and any replication attempts of suitable quality.

Unlike full papers elsewhere, these *Brief Reports* would consist of the method and results section only. This greatly reduces the cost of either writing or reviewing the report. The Brief Report must also be submitted with one or more RepLinks, specifying what exactly is being replicated. Particularly for non-replications, authors of Brief Reports can use the comments on the RepLinks to discuss why they think the replication failed (low power in the original study, etc.).

Review of Brief Reports would be handled by moderators. When a Brief Report is submitted, all moderators of that sub-field would be automatically emailed with a request to review the proposed post. The review could then be ‘claimed’ by any moderator. If no one claims the post for review within a week, the system would then automatically choose one of the relevant moderators, and ask if they would accept the request to review; if they decline, further requests would be made until someone agreed to review. Authors would not be able to be the

sole moderator/reviewer for replications of their own work. As in the PLoS model, the moderator could evaluate the *Brief Report* alone or solicit outside review(s).

The presumption of the review process would be acceptance. Brief Reports would be returned for revision when appropriate, as in the case of using inappropriate statistical tests; but would only be rejected if the paper does not actually qualify as a replication attempt (based on the criteria discussed above). In the latter case, authors of Brief Reports could appeal the decision, which would then be reviewed by two other moderators. On acceptance, the Brief Report would be published online in static form with a DOI, much like any other publication, and thus be part of the citable, peer-reviewed record. The appropriate RepLinks would be likewise added to Replication Tracker. As with any RepLink, these could be suppressed if flagged as irrelevant a sufficient number of times (see above). Thus, while publication in Brief Reports is permanent (barring retractions), incorporation into Replication Tracker is always potentially in flux -- as is appropriate for a post-review evaluation process.

The experience of using Replication Tracker: A step-by-step guide

As in any literature database, users would begin by using a search function (either simple or advanced) to locate a paper of interest (Figure 1). This search would bring up a list of references, in a format similar to Google Scholar. However, in addition to the citation count provided by Google Scholar, the system would provide three additional values: The number of replication attempts documented, the paper's Replicability Score, and the Strength of Evidence score (Figure 2). As described above, the Replicability Score would hold a value from -2 to +2, with negative values denoting evidence of non-replication, zero denoting mixed findings, and positive values evidence of successful replication.

The user would then click on a reference from the list to bring up more detailed information about that target paper (Figure 3). The target paper's reference would appear at the top of the page, along with the number of attempted replications documented, Replicability Score for that paper, and the Strength of Evidence score. Below these aggregate measures would be a list of the RepLinks, represented by a citation of the RepLinked paper, the aggregate Type of Finding score and Strength of Evidence score for that RepLink, and the number of users who have rated that RepLink. An additional button would allow users to add their own ratings or flag the RepLink as irrelevant.

Information about each RepLink could be expanded, to show each individual rating along with that users' associated comments, if any (Figure 4). Users could agree with an existing rating via a thumbs-up button. Ratings and comments would be labeled with the username of the poster; for authenticated accounts, they could optionally be labeled with the individuals' real name. Comments by authors who have chosen to authenticate their account under their real names would be labeled as such.

Issues for Further Discussion

The Replication Tracker would serve several functions. First, it would enable a new way of navigating the literature. Second, we believe it would motivate researchers to conduct and report attempted replications, helping correct biases in the literature such as the file-drawer problem. Third, it will vastly improve access to and communication regarding replication attempts. Perhaps most importantly, it would help incentivize and reward costly efforts to ensure replicability pre-publication, helping to mitigate a Tragedy of the Commons in scientific publishing.

However, in addition to these potential benefits, tracking and publishing replication attempts raises non-trivial issues, and has the potential for unintended consequences. We consider several such concerns below and discuss how these concerns may be addressed or allayed.

Getting the system off the ground

The usefulness of the database for tracking replicability will be a function of the amount of replication information added to it, in the form of RepLinks, metadata information, and Brief Reports. This will require considerable participation by a broad swath of the research community. Because researchers are more likely to contribute to a system that they already find useful, an important determiner of success will be the ability to achieve a critical mass of such information. We have considered several ways of increasing the likelihood that the system quickly reaches critical mass.

First, there should be a considerable number of founding members, so that a wide range of researchers are engaged in the project prior to launch. This will not only help with division of labor, but will also help clarify the many design decisions that go into creating the details of the system. The more diverse the founding group is, the more likely the final system will be acceptable to researchers in multiple fields and disciplines. This paper serves as a first step in starting the needed dialog.

Second, we suggest concentrating on first reaching critical mass for a few select subfields of psychology and neuroscience, instead of simultaneously attempting to obtain critical mass in all fields of science at once. In order to reach critical mass within the first few subfields, we suggest that prior to the public launch of Replication Tracker, founding members conduct targeted replicability reviews of specific literatures within those subfields, writing RepLinks and

soliciting Brief Reports during the process. These data would be used to write review papers, which would be published in traditional journals. These review papers would be useful publications in and of themselves, and would help demonstrate the empirical value of tracking replications. This would help recruit additional founders, moderators and funding -- all while major components are added to the database. Only once enough coverage of the literatures within those subfields has been achieved would Replication Tracker be publically launched.

In addition to tracking published replications, the proposed system attempts to ameliorate the file-drawer problem by allowing researchers to submit Brief Reports of attempted replications. Several previous attempts have been made to publish null results and replication attempts (e.g. Journal of Articles in Support of the Null Hypothesis; Journal of Negative Results in Biomedicine) often with low rates of participation (JASNH has published 32 papers since its launch in 2002). Nonetheless, we believe several aspects of our system would motivate increased participation. Firstly, the format of Brief Reports significantly decreases the time commitment of preparation, as the Reports consist of the method and results section only. Second, these Brief Reports will not only be citable, but will also be highly findable, as they will be RepLinked to the relevant published papers. Thus we expect these Reports to have some value, perhaps equivalent to a conference paper or poster. We believe that the combination of lesser time investment and increased value will lead to increased rates of submission.

What is the right unit of analysis?

Because each paper may include multiple findings that differ in replicability, there is a good argument to be made that what should be tracked is the replicability of a given result. We propose tracking the replicability of papers instead, for several reasons.

The first reason is one of feasibility. We believe that tracking each finding separately would be infeasible, as what counts as an individual finding may be subjective, and the vast number of units of analysis even within a single paper becomes prohibitive. An intermediate level would be to track individual experiments. However, publication formats do not always include separate headings for each individual experiment (e.g. *Nature*, *Current Biology*), and even a single experiment may include multiple components with differences in replicability.

Secondly, even organizing the system at the level of experiment will not allow an aggregated replicability score to capture every nuance of the scientific literature. It will always be necessary for the reader to examine written information for more detail, including the full text of the RepLinked papers. For these detail-oriented readers, the proposed system provides a novel way to navigate through published work (by following RepLinks to find and read papers with attempted replications) and an efficient way to view comments on each of these papers (Figure 4). Such a system is most intuitive and navigable when organized at the level of the paper itself.

Are sufficient numbers of replications conducted?

The rate of published replications appears to be low: For instance, over a 20-year period, only 5.3% of 701 publications in nine management journals included attempts to replicate previous findings (Hubbard, Vetter, & Little, 1998). While we believe Replication Tracker would lead to increased numbers of published replications, we must consider whether Replication Tracker would be useful if the number of published replications remains low. Certainly, many papers will simply never be replicated, and many others will only have one reported replication attempt.

We do not believe these issues undermine the utility of Replication Tracker for several reasons. First, the findings which are of broadest interest to the community are likely the very

same findings for which the most replications are attempted. Thus, while many low-impact papers may lack replication data, the system will be most useful for the papers where it is most needed. Secondly, even low numbers of replications are often sufficient: because spurious results are unlikely to replicate, even only a handful of successful replications significantly increases the likelihood that a given finding is real (Moonesinghe, Khoury, & Janssens, 2007). Finally, we note that even sparse replicability data is useful when aggregating over large numbers of papers, for instance, when producing aggregate Replicability Scores for journals. Similarly, it would be possible to aggregate across studies within individual literatures or using particular methods. For these aggregate scores, sparse data does not present a problem.

Would tracking replicability stifle novel scientific fields?

Commenters on the present paper have suggested that since new fields may still be designing the details of their methods, and may be less sure of what aspects of the method are necessary to correctly measure the effects under investigation, their initial results may appear less replicable. In this case, using replicability scores as a measure of paper, researcher and journal quality -- one of our explicit aims -- could potentially stifle new fields of enquiry.

This is an important concern if true. We do not know of any systematic empirical data that would adjudicate the issue. However, we suspect that other factors may systematically increase replicability in new lines of inquiry. For example, young fields may focus on larger effects, with established fields focusing on increasingly subtle effects over time (cf Taubes & Mann, 1995). Additionally, in the case that subtle methodological differences prevent replication of results, Replication Tracker may actually aid researchers in identifying the relevant issues more quickly, spurring growth of the novel field.

We additionally note that it is not our intention that replicability become the sole criteria by which research quality is measured, nor do we think that is likely to happen. New fields are likely to generate excitement and citations, which will produce their own momentum. The goal is that replicability rates be considered in addition.

Would Replication Tracker underestimate replicability?

Commenters on the present paper have also suggested several ways in which Replication Tracker might underestimate replicability. Underestimating the replicability of a field could undermine both scientists' and the public's confidence in the field, leading to decreased interest and funding.

Null effect bias. Researchers may be more motivated to submit non-replications to the system as Brief Reports, while successful replications would languish in file-drawers. We suspect that this problem would disappear as the system gains popularity: Researchers typically attempt replications of effects that are crucial to their own line of work, and will find it useful to report those replications in order to have their own work embedded in a well-supported framework. Moreover, many replication attempts are conducted by the authors of the original study, who will be intrinsically motivated to report successful replications in support of their own work. Nonetheless, this is an issue that should be evaluated and monitored as Replication Tracker is introduced, so that adjustments can be made as necessary.

Unskilled replicators. Another concern is that if on average, the researchers that tend to conduct large numbers of strict replications are less skilled than the original researchers, this could lead to non-replications due to unknown errors. If this is the case, this issue could be compensated for in two ways. First, as Replication Tracker and Brief Reports raise the profile of replication, more skilled researchers may begin to conduct and report more replications. Second,

as discussed above, there are numerous machine learning techniques to identify the most reliable sources of information. These techniques could be applied to mitigate this issue, by discounting replication data from users that have not been reliable sources of information in the past.

Spurious non-replications. Since the statistical power to detect an effect is never 1.0, even true effects sometimes do not replicate. High-profile papers in particular will be much more likely to be subject to replication attempts; since some replications even of real effects will fail, high-profile papers may be unfairly denigrated. This issue is compounded if typical statistical power in that literature is low, making replication improbable.

These issues can be dealt with directly in Replication Tracker, by appropriately weighing this probabilistic information. Recall that Replication Tracker provides both a Replicability Score, indicating whether existing evidence suggests that the target paper replicates, as well as a Strength of Evidence Score. A single non-replication -- particularly one with only mid-sized power -- is not strong evidence for non-replicability, and this should be reflected in the Strength Score. Replication attempts with low power should not be RepLinked at all. If 8 of 10 replication attempts succeed -- consistent with statistical power of .8 -- that should be counted as strong evidence of replicability.

Will Type II error increase?

Finally, we must consider whether the changes people will make to their work will actually lead to an increased d' (ability to detect true effects) or whether these changes will simply result in a tradeoff: researchers may eliminate some false positives (Type I error) only at the expense of increasing the false negative rate (Type II error). It is an open question whether fields like psychology and neuroscience are currently at an optimal balance between Type I and Type II error, and Replication Tracker would help provide data to adjudicate this issue.

Moreover, some of the potential reforms would almost certainly increase d' , like conducting studies with greater statistical power.

Conclusion

In conclusion, we propose tracking replication attempts as a key method of identifying high-quality research post-publication. We argue that tracking and incentivizing replicability directly would allow researchers to escape the current Tragedy of the Commons in scientific publishing, thus helping to speed the adoption of reforms. In addition, by tracking replicability, we will be able to determine whether any adopted reforms have successfully increased replicability.

No measure of research quality can be perfect; instead, we aim to create a measure that is robust enough to be useful. Citation counts have proven very useful in spite of the metrics' many flaws as measure of a paper's quality (for instance, papers which are widely criticized in subsequent literature will be highly cited). Tracking replicability and tracking citations have complementary strengths and weaknesses: Influential results may not be replicable. Replicable results may not be influential. The combination of both metrics should allow us to identify results that are both influential and replicable, thus more accurately identifying high-quality empirical work.

References

- Adamic, L. A., Zhang, J., Bakshy, E., & Ackerman, M. S. (2008). Knowledge sharing and yahoo answers: Everyone knows something. *Proceedings of the 17th International Conference on World Wide Web*.
- Albert, P. S., & Dodd, L. E. (2004). A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics*, *60*, 427-435.
- Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated Significance Tests on Accumulating Data. *Journal of the Royal Statistical Society: Series A*, *132*, 235-244.
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390-412.
- Bederson, B. B., Hu, C., & Resnik, P. (2010). Translation by interactive collaboration between monolingual users. *Proceedings of Graphics Interface*, 39-46.
- Bezeau, S., & Graves, R. (2001). Statistical power and effect sizes of clinical neuropsychology research. *Journal of Clinical and Experimental Neuropsychology*, *2001*(23), 3.
- Boffetta, P., Mclaughlin, J. K., Vecchia, C. L., Tarone, R. E., Lipworth, L., & Blot, W. J. (2008). False-Positive Results in Cancer Epidemiology : A Plea for Epistemological Modesty. *Journal of the National Cancer Institute*, *100*, 988-995.
- Callahan, M. L., Wears, R. L., Weber, E. J., Barton, C., & Young, G. (1998). Positive-outcome bias and other limitations in the outcome of research abstracts submitted to a scientific meeting. *JAMA : the journal of the American Medical Association*, *280*, 254-257.
- Callison-Burch, C. (2009). Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 286-295.
- Chase, L. J., & Chase, R. B. (1976). A statistical power analysis of applied psychological research. *Journal of Applied Psychology*, *61*(2), 234-237.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research. *Journal of Abnormal and Social Psychology*, *65*, 145-153.
- Cole Jr., S., Cole, J. R., & Simon, G. A. (1981). Chance and consensus in peer review. *Science*, *214*, 881-886.
- Coursol, A., & Wagner, E. E. (1986). Effect of positive findings on submission and acceptance rates: A note on meta-analysis bias. *Professional Psychology: Research and Practice*, *17*, 136-137.
- Cuijpers, P., Smit, F., Bohlmeijer, E., Hollon, S. D., & Andersson, G. (2010). Efficacy of cognitive-behavioral therapy and other psychological treatments for adult depression: Meta-analytic study of publication bias. *The British Journal of Psychiatry*, *196*, 173-178.
- Dewald, W. G., Thursby, J. G., & Anderson, R. G. (1986). Replication in Empirical Economics : The Journal of Money , Credit and Banking Project. *The American Economic Review*, *76*, 587-603.
- Dickersin, K., Chan, S., Chalmers, T. C., Sacks, H. S., & Smith, H. (1987). Publication bias and clinical trials. *Controlled clinical trials*, *8*, 343-353.
- Dickersin, K., Min, Y.-I., & Meinert, C. L. (1992). Factors influencing publication of research results: Follow-up of applications submitted to two institutional review boards. *Journal of the American Medical Association*, *267*, 374-378.
- Doan, A., Ramakrishnan, R., & Halevy, A. Y. (2011). Crowdsourcing systems on the World-Wide Web. *Communications of the ACM*, *54*(4), 86-96.
- Dwan, K., Altman, D. G., Arnaiz, J. a., Bloom, J., Chan, A.-W., Cronin, E., et al. (2008). Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS one*, *3*, e3081.

- Easterbrook, P. J., Berlin, J. A., Gopalan, R., & Matthews, D. R. (1991). Publication bias in clinical research. *The Lancet*, 337, 867-872.
- Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193-242.
- Ernst, C., & Angst, J. (1983). *Birth Order: Its Influence on Personality*. New York: Springer-Verlag.
- Eysenck, H. J., & Eysenck, S. B. (1992). Peer review: Advice to referees and contributors. *Personality and Individual Differences*, 13(4), 393-399.
- Feller, W. (1940). Statistical aspects of ESP. *Journal of Parapsychology*, 4, 271-298.
- Ferguson, C. J., & Kilburn, J. (2010). Much Ado About Nothing: The Misestimation and Overinterpretation of Violent Video Game Effects in Eastern and Western Nations: Comment on Anderson et al. (2010). *Psychological Bulletin*, 5-9.
- Field, M., Munafo, M. R., & Franken, I. H. A. (2009). A meta-analytic investigation of the relationship between attentional bias and subjective craving in substance abuse. *Psychological Bulletin*, 135(4), 589-607.
- Flint, J., & Munafo, M. R. (2009). Replication and heterogeneity in gene x environment interaction studies. *International Journal of Neuropsychopharmacology*, 727-729.
- Franklin, M., Kossmann, D., Kraska, T., Ramesh, S., & Xin, R. (2011). *CrowdDB: Answering queries with crowdsourcing*. Paper presented at the SIGMOD 2011.
- Gehr, B. T., Weiss, C., & Porzolt, F. (2006). The fading of reported effectiveness. A meta-analysis of randomised controlled trials. *BMC medical research methodology*, 6, 25.
- Gerberg, A. S., Green, D. P., & Nickerson, D. (2001). Testing for publication bias in political science. *Political Analysis*, 9(4), 385-392.
- Giles (2005). Internet encyclopedias go head to head. *Nature*, 438, 900-901.
- Hardin, G. (1968). The Tragedy of the Commons. *Science*, 162(3859), 1243-1248.
- Harris, J. R. (1998). *The Nurture Assumption: Why Children Turn out the Way that They Do*. New York: Free Press.
- Hartshorne, J. K., Salem-Hartshorne, N., & Hartshorne, T. S. (2009). Birth order effects in the formation of long-term relationships. *Journal of Individual Psychology*, 65(2), 156-176.
- Hirschhorn, J. N., Lohmueller, K. E., Byrne, E., & Hirschhorn, K. (2002). A comprehensive review of genetic association studies. *Genetics in Medicine*, 4, 45-61.
- Hubbard, R., Vetter, D. E., & Little, E. L. (1998). Replication in strategic management: scientific testing for validity, generalizability, and usefulness. *Strategic Management Journal*, 19, 243-254.
- Ioannidis, J. P. A. (2005a). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association*, 294(2), 218-228.
- Ioannidis, J. P. A. (2005b). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.
- Ioannidis, J. P. A., Ntzani, E. E., Trikalinos, T. A., & Contopoulos-Ioannidis, D. G. (2001). Replication validity of genetic association studies. *Nature genetics*, 29, 306-309.
- Ioannidis, J. P. A., Trikalinos, T. A., Ntzani, E. E., & Contopoulos-Ioannidis, D. G. (2003). Genetic associations in large versus small studies: An empirical assessment. *The Lancet*, 361(9357), 567-571.
- Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59, 434-446.
- Jennions, M. D., & Møller, A. P. (2002a). Publication bias in ecology and evolution: an empirical assessment using the 'trim and fill' method. *Biological reviews of the Cambridge Philosophical Society*, 77, 211-222.
- Jennions, M. D., & Møller, A. P. (2002b). Relationships fade with time: a meta-analysis of temporal trends in publication in ecology and evolution. *Proceedings. Biological sciences / The Royal Society*, 269, 43-48.

- Kosciulek, J. F., & Szymanski, E. M. (1993). Statistical power analysis of rehabilitation counseling research. *Rehabilitation Counseling Bulletin*, 36(4), 212-219.
- Kristensen, P., & Bjerkedal, T. (2007). Explaining the relation between birth order and intelligence. *Science*, 316, 1717.
- Law, E., von Ahn, L., Dannenberg, R., & Crawford, M. (2007). *TagATune: A game for sound and music annotation*. Paper presented at the ISMIR.
- Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S., & Hirschhorn, J. N. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature genetics*, 33, 177-182.
- Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, 1(2), 161-175.
- Misakian, A. L., & Bero, L. A. (1998). On passive smoking: Comparison of published and unpublished studies. *Journal of the American Medical Association*, 280, 250-253.
- Mone, M. A., Mueller, G. C., & Mauland, W. (1996). The perceptions and usage of statistical power in applied psychology and management research. *Personnel Psychology*, 49, 103-120.
- Moonesinghe, R., Khoury, M. J., & Janssens, a. C. J. W. (2007). Most published research findings are false-but a little replication goes a long way. *PLoS medicine*, 4, e28.
- Munafò, M. R., Stothart, G., & Flint, J. (2009). Bias in genetic association studies and impact factor. *Molecular psychiatry*, 14, 119-120.
- Newton, D. P. (2010). Quality and peer review of research: an adjudicating role for editors. *Accountability in research*, 17, 130-145.
- Olson, C. M., Rennie, D., Cook, D., Dickersin, K., Flanagan, A., Hogan, J. W., et al. (2002). Publication bias in editorial decision making. *JAMA : the journal of the American Medical Association*, 287, 2825-2828.
- Peters, D. P., & Ceci, S. J. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences*, 5, 187-195.
- Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology*, 7, 331-363.
- Saxe, R., Brett, M., & Kanwisher, N. (2006). Divide and conquer: A defense of functional localizers. *NeuroImage*, 30, 1088-1096.
- Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effects on the power of studies? *Psychological Bulletin*, 105(2), 309-316.
- Sena, E. S., Worp, H. B. V. D., Bath, P. M. W., Howells, D. W., & Macleod, M. R. (2010). Publication Bias in Reports of Animal Stroke Studies Leads to Major Overstatement of Efficacy. *Review Literature And Arts Of The Americas*, 8.
- Shen, W., Kiger, T. B., Davies, S. E., Rasch, R. L., Simon, K. M., & Ones, D. S. (in press). Samples in applied psychology: Over a decade of research in a review. *Journal of Applied Psychology*.
- Snow, R., O'Conner, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast - but is it good? Evaluating non-expert annotations for natural language tasks. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 254-263.
- Taubes, G., & Mann, C. C. (1995). Epidemiology faces its limits. *Science*, 269(5221), 164-169.
- Trikalinos, T. A., Ntzani, E. E., Contopoulos-Ioannidis, D. G., & Ioannidis, J. P. A. (2004). Establishment of genetic associations for complex diseases is independent of early study findings. *European Journal of Human Genetics*, 12, 762-769.
- Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76(2), 105-110.
- von Ahn, L., & Dabbish, L. (2008). General techniques for designing games with a purpose. *Communications of the ACM*, 51(8), 58-67.

- von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). reCAPTCHA: Human-based character recognition via web security measures. *Science*, 321(5895), 1465-1468.
- Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspectives on Psychological Science*, 4, 274-290.
- Wagenmakers, E. J., Wetzels, R., Borsboom, D., & Van, H. (in press). Why Psychologists Must Change the Way They Analyze Their Data : The Case of Psi. *Psychology*, 1-14.
- Wager, T. D., Lindquist, M., & Kaplan, L. (2007). Meta-analysis of functional neuroimaging data: Current and future directions. *Social Cognitive and Affective Neuroscience*, 2(2), 150-158.
- Wager, T. D., Lindquist, M. A., Nichols, T. E., Kober, H., & van Snellenberg, J. X. (2009). Evaluating the consistency and specificity of neuroimaging data using meta-analysis. *NeuroImage*, 45, S210-S221.
- Welinder, P., Branson, S., Belongie, S., & Perona, P. (2010). *The multidimensional wisdom of the crowds*. Paper presented at the Advances in Neural Information Processing Systems 2010.
- Yan, T., Kumar, V., & Ganesan, D. (2010). CrowdSearch: Exploiting crowds for accurate real-time image search on mobile phones. *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*.
- Yarkoni, T. (2009). Big Correlations in Little Studies: Inflated fMRI Correlations Reflect Low Statistical Power-Commentary on Vul et al. (2009). *Perspectives on Psychological Science*, 4, 294-298.
- Yarkoni, T., & Braver, T. S. (2010). Cognitive neuroscience approaches to individual differences in working memory and executive control: Conceptual and methodological issues. In A. Gruzka, G. Matthews & B. Szymura (Eds.), *Handbook of Individual Differences in Cognition: Attention, Memory, and Executive Control* (pp. 87-108). New York: Springer.

Appendix: Survey Methods and Results

We contacted 100 colleagues directly as part of an anonymous Web-based survey. Colleagues of the authors from different institutions were invited to participate, as well as the entire faculty of one research university and one liberal-arts college. 49 individuals completed the survey: 26 faculty members, 9 post-docs, and 14 graduate students. 38 of these participants worked at national research universities. Respondents represented a wide range of sub-disciplines: clinical psychology (2), cognitive psychology (11), cognitive neuroscience (5), developmental psychology (10), social psychology (6), school psychology (2), and various inter-subdisciplinary areas.

The survey was presented using Google Survey. Participants filled out the survey at their leisure during a single session. The full text of the survey, along with summaries of the results, is included below. All research was approved by the Harvard University Committee on the Use of Human Subjects, and informed consent was obtained.

Part 1: Demographics

Your research position: graduate student, post-doc, faculty, other (26 faculty, 9 post-docs, and 14 graduate students).

Your institution: national university, regional university, small liberal arts college, other (38 national university, 4 regional university, 5 small liberal arts college, 2 other).

Your subfield (cognitive, social, developmental, etc.; There is no standard set of subfields. Use your own favorite label): _____

(11 cognitive psychology, 10 developmental psychology, 6 social psychology, 5 cognitive neuroscience, 2 school psychology, 2 clinical psychology, 13 multiple/other).

Part 2: Completed Replications

In this section, you will be asked about your attempts to replicate published findings. When we say “replication”, we mean:

-a study in which the methods are designed to be as similar as possible to a previously published study. There may be minor differences in the method so long as they are not expected to matter under any existing theory. However, a study which uses a different method to make a similar or convergent theoretical point would be more than a replication. If you attempted to replicate the same finding several times, each attempt should be counted separately.

Given this definition...

1) Approximately how many times have you attempted to replicate a published study? Please count only completed attempts -- that is, those with at least as many subjects as the original study. _____

Total: 257; Mean: 6; Median: 2; SD: 11

(3 excluded: "NA", "too many to count", "50+")

*2) How many of these attempts ***fully*** replicated the original findings? _____*

Excluding those excluded in (1):

Total: 127; Mean: 4; Median: 1; SD: 7

*3) How many of these attempts ***partially*** replicated the original findings? _____*

Excluding those excluded in (1):

Total: 77; Mean: 2; Median: 1; SD: 5

4) How many of these attempts failed to replicate any of the original findings? _____

Excluding those excluded in (1):

Total: 79; Mean: 2; Median: 1; SD: 4

5) Please add any comments about this section here: _____

[comments]

Part 3: Aborted Replications

In this section, you will be asked about attempted replications that you did not complete (e.g., tested fewer participants than were tested in the original study).

1) *Approximately how many times have you started an attempted replication but stopped before collecting data from a full sample of participants? _____*

Total: 48; Mean: 1; Median: 0; SD: 3

(3 excluded: "a few", "countless", [lengthy discussion])

2) *Of these attempts, how many were stopped because the data thus far failed to replicate the original findings? _____*

Excluding those excluded in (1):

Total: 38; Mean: 2; Median: 0.5; SD = 4

3) *Of these attempts, how many were stopped for another reasons (please explain)? _____*

[comments]

4) *Please add any comments about this section here.*

[comments]

Part 4: File Drawers

1) *Approximately how many experiments have you completed (collected the full dataset) but, at this point, do not expect to publish? _____*

Total: 1312 (one participant reported "1000"); Mean: 31; Median: 3.5; SD: 154

(6 excluded: "many", "ton", "countless", "30%-50%?", 2 unreadable/corrupted responses)

2) Of these, how many are not being published because they did not obtain any statistically significant findings (that is, they were null results)? ____

Excluding those excluded in (1):

Total: 656 (one participant reported "500"); Mean: 17; Median: 2; SD: 81

3) Please add any comments about this section here: ____

[comments]

Figure Captions

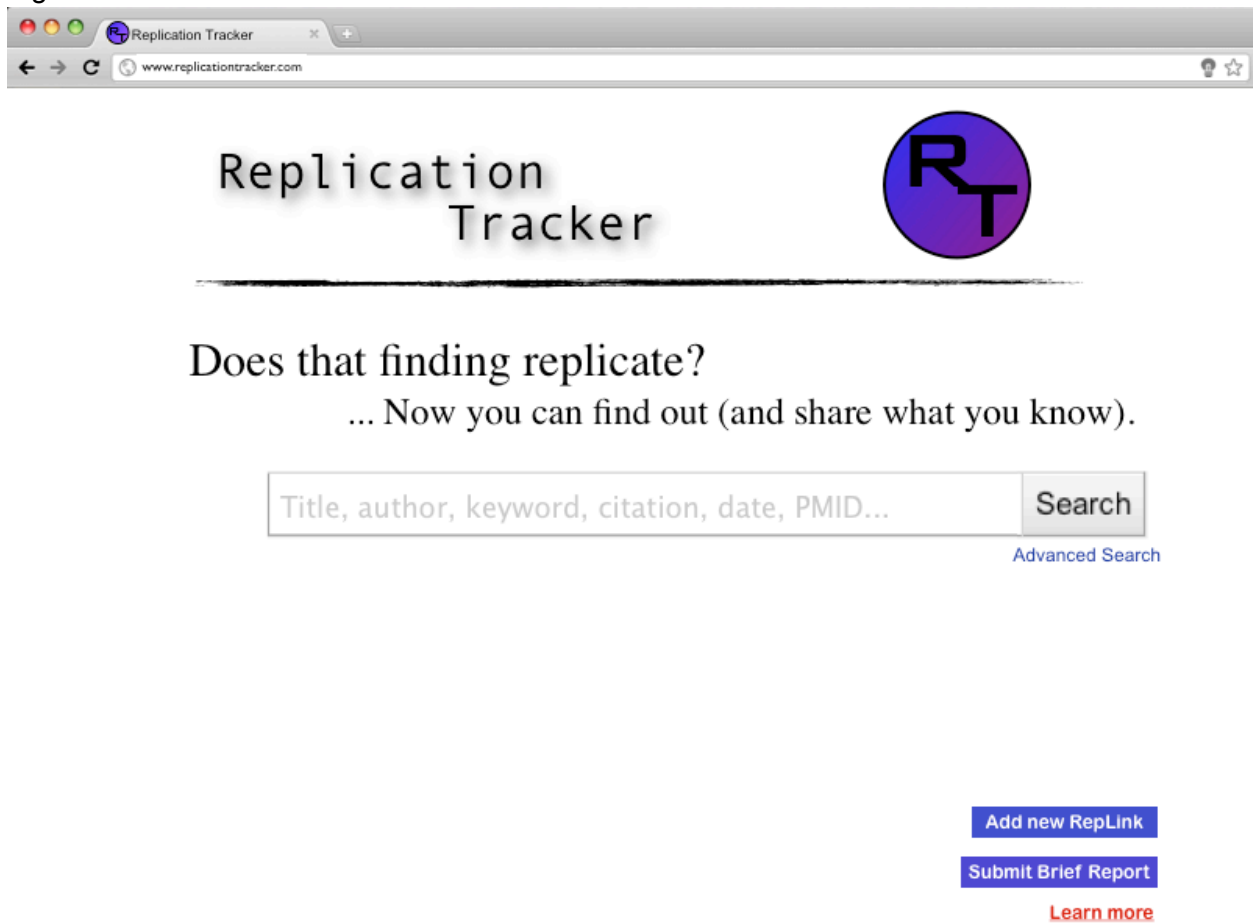
Figure 1. Replication Tracker: Search window.

Figure 2. Replication Tracker: Example search results.

Figure 3. Replication Tracker: Search results expansion, showing RepLinks for a target paper.

Figure 4. Replication Tracker: Expansion of a RepLink, showing ratings by individual readers.

Figure 1.



The image shows a screenshot of a web browser displaying the homepage of the Replication Tracker website. The browser's address bar shows the URL www.replicationtracker.com. The page features the site's logo, which consists of the letters 'RT' in a stylized font inside a purple circle. Below the logo, the text 'Replication Tracker' is displayed in a large, serif font. A horizontal line separates the header from the main content area. The main content area contains the text 'Does that finding replicate?' followed by the subtitle '... Now you can find out (and share what you know)'. Below this text is a search bar with the placeholder text 'Title, author, keyword, citation, date, PMID...' and a 'Search' button. To the right of the search bar is a link for 'Advanced Search'. At the bottom right of the page, there are three buttons: 'Add new RepLink', 'Submit Brief Report', and 'Learn more'.

Replication Tracker

Does that finding replicate?
... Now you can find out (and share what you know).

Title, author, keyword, citation, date, PMID... Search

[Advanced Search](#)

[Add new RepLink](#)
[Submit Brief Report](#)
[Learn more](#)

Figure 2.

The screenshot shows a web browser window with the URL www.replicationtracker.com. The page features the Replication Tracker logo (a purple circle with 'RT' in white) and a search bar containing the text 'working memory capacity'. To the right of the search bar are buttons for 'Search', 'Add new RepLink', and 'Submit Brief Report', along with a link to 'Advanced Search'. Below the search bar, a horizontal line separates the header from the search results. The results are listed as follows:

- [The capacity of working memory: What are the limits?](#)
JQ Sample, IA Author - Trends in Cognitive Sciences - 2005
Cited by 1082 - Replication Attempts: 8 - Replicability Score: 1.3 (partial replication) - Evidence: 4 (Strong)
- [The role of statistical regularities in visual working memory](#)
IA Author, JQ Sample - JEP: General - 1999
Cited by 326 - Replication Attempts: 3 - Replicability Score: 2.5 (full replication) - Evidence: 2 (Weak)
- [Working memory capacity for real-world objects](#)
ME Sample, UA Author - Psychological Science - 2008
Cited by 12 - Replication Attempts: 1 - Replicability Score: 1 (partial replication) - Evidence: 1 (Weak)
- [Working memory capacity across the lifespan](#)
IA Author, JQ Sample - Journal of Aging- 2001
Cited by 5 - Replication Attempts: 0
- [The effect of musical training on working memory capacity](#)
ME Sample, UA Author - Music Perception - 2011
Cited by 1 - Replication Attempts: 0

At the bottom right of the search results, there is a blue link labeled [Next](#).

Figure 3.

The screenshot shows a web browser window with the URL www.replicationtracker.com. The page header includes the "Replication Tracker" logo, a search bar containing the text "working memory capacity", and buttons for "Add new RepLink", "Submit Brief Report", and "Advanced Search".

The search results display the following information:

- [The capacity of working memory: What are the limits?](#)
JQ Sample, IA Author - Trends in Cognitive Sciences - 2005
Cited by 1082 - Replication Attempts: 8 - Replicability Score: 1 (partial replication) - Evidence: 4 (Strong)

Two result entries are highlighted with red dashed circles and labeled "Rate it!":

- Rate it!** Type of Finding: 1 (Partial replication) Strength of Evidence: 4 (Strong) # of Ratings: 3
[Smith & Shmoe, 2011. Replicability Tracker Brief Report \(read fulltext\)](#)
Expand for details and comments
- Rate it!** Type of Finding: 0 (Mixed) Strength of Evidence: 1 (Weak) # of Ratings: 2
[Sample & Author, 2008. Cognition. \(read fulltext\)](#)
Expand for details and comments

Figure 4.

The screenshot shows the Replication Tracker website interface. At the top, there is a search bar containing the text "working memory capacity" and a "Search" button. Below the search bar are two buttons: "Add new RepLink" and "Submit Brief Report". To the right of these buttons is a link for "Advanced Search".

The main content area displays the search results for "working memory capacity". The top result is titled "The capacity of working memory: What are the limits?" by "JQ Sample, IA Author" from "Trends in Cognitive Sciences - 2005". Below the title, it shows citation statistics: "Cited by 1082 - Replication Attempts: 8 - Replicability Score: 1 (partial replication) - Evidence: 4 (Strong)".

A detailed view of a replication attempt is shown in a rounded rectangular box. On the left side of this box is a red circular button with the text "Rate it!". The detailed view includes the following information:

- Type of Finding:** 1 (Partial replication)
- Strength of Evidence:** 4 (Strong)
- # of Ratings:** 3

Below this summary are three individual ratings from users:

 JQ Sample - authenticated user - author	"Smith & Shmoe present interesting findings. We wish to point out that their method differed from ours, in the duration of each test trial."
 MsProf - authenticated user	"This is a clean replication."
 visionScientist	(no comment provided)

At the bottom of the detailed view box is a button labeled "Close details and comments".