DEPARTMENT OF ECONOMICS AND FINANCE
SCHOOL OF BUSINESS
UNIVERSITY OF CANTERBURY
CHRISTCHURCH, NEW ZEALAND

**A Primer on the “Reproducibility Crisis” and
Ways to Fix It**

W. Robert Reed

WORKING PAPER

No. 21/2017

**Department of Economics and Finance
School of Business
University of Canterbury
Private Bag 4800, Christchurch
New Zealand**

WORKING PAPER No. 21/2017

A Primer on the “Reproducibility Crisis” and Ways to Fix It

W. Robert Reed^{1†}

December 2017

Abstract: This article uses the framework of Ioannidis (2005) to organize a discussion of issues related to the “reproducibility crisis.” It then goes on to use that framework to evaluate various proposals to fix the problem. Of particular interest is the “post-study probability”, the probability that a reported research finding represents a true relationship. This probability is inherently unknowable. However, a number of insightful results emerge if we are willing to make some conjectures about reasonable parameter values. Among other things, this analysis demonstrates the important role that replication can play in improving the signal value of empirical research.

Keywords: Reproducibility crisis, Post-study probability, Significance level, Power, Publication bias, Pre-registration, Registered reports, Negative results, Replication

JEL Classifications: A12, B41, C10, C80

Acknowledgements: We gratefully acknowledge comments and feedback from Tom Coupé, Brian Haig, Daniël Lakens, Andrea Menclova, E.J. Wagenmakers, and Benjamin D.K. Wood. Their acknowledgement should not be interpreted to mean that they agree with everything stated in this article.

¹ Department of Economics and Finance, University of Canterbury, NEW ZEALAND

† The corresponding author is W. Robert Reed. Email: bob.reed@canterbury.ac.nz.

I. Introduction

The last two decades have seen increasing doubt about the credibility of empirical research in science. This has come to be known as the “Reproducibility Crisis,” with the name derived from the fact that many reported empirical findings cannot be reproduced/replicated. While concerns have been raised about all areas of science, medicine and the social sciences, particularly psychology, have been the subject of greatest concern. Economics has been relatively slow to recognize the problem and consider solutions.

A thorough discussion of the evidence for a reproducibility crisis would require considerable space, and thus cannot be undertaken here. Suffice it to say that while an increasing number of researchers are convinced that there is a problem, others remain unpersuaded. The 2017 *Papers and Proceedings* issue of the *American Economic Review* provides a sampling of different perspectives.¹ This article will adopt the framework developed by Ioannidis (2005) to organize a discussion of issues related to the “reproducibility crisis”, and to analyze some of the solutions that have been proposed to fix it. It will devote special attention to replication.

II. The logic behind the reproducibility crisis

In 2005, John Ioannidis published a paper entitled “Why Most Published Research Findings are False” (Ioannidis, 2005). It is among the most highly cited papers in medicine and the social sciences. In the paper, he presents a very simple mathematical model to show that it is highly unlikely that any published paper that claims a statistically significant relationship is actually true.

In its simplest form, the model consists of three components. The first component is π , the probability that the relationship being studied actually exists. In a regression context, where

¹ See particularly the papers in the sessions (i) “Replication in Microeconomics,” and (ii) “Replication and Ethics in Economics: Thirty Years After Dewald, Thursby, and Anderson” (*American Economic Review*, 2017).

one is estimating $y_t = \varphi_0 + \varphi_1 x_t + \varepsilon_t$, π is the probability that $\varphi_1 \neq 0$. In any given study, π is either 1 or 0. But consider a large number of published studies, all exploring possible relationships between different y 's and x 's. Some of these relationships will really exist in the population, and some will not. π is the probability that a randomly chosen published study is estimating a relationship that truly exists in the population.

The second component is α , the probability of a Type I error. In our context, it is the probability of rejecting the hypothesis $H_0: \varphi_1 = 0$ when it is true. Assuming a model is correctly specified, a researcher can control α by setting the appropriate critical value for the relevant sample statistic. α is commonly set at 0.05 in the social sciences.

The third component is the power of the study, defined as $(1 - \beta)$, where β is the probability of a Type II error. In our context, if β is the probability of failing to reject $H_0: \varphi_1 = 0$ when it is false, then $(1 - \beta)$ is the probability of correctly rejecting it. Of course, unlike α , which is a constant set by the researcher, $(1 - \beta)$ will vary depending on a number of factors, such as the size of the true effect and the variance of the error term. Despite the fact that it is very difficult to measure the power of a study in an actual research situation, researchers have made educated guesses about likely values for $(1 - \beta)$ in the social sciences.

With these three components, we are in a position to calculate four probabilities:

- P_1 : the probability that a published study reports a statistically significant estimate and a relationship truly exists;
- P_2 : the probability that a published study reports a statistically insignificant estimate and a relationship truly exists;
- P_3 : the probability that a published study reports a statistically significant estimate and a relationship does not exist;
- P_4 : the probability that a published study reports a statistically insignificant estimate and a relationship does not exist.

Table 1 reports each of these probabilities as a function of π , α , and $(1 - \beta)$. For example, P_1 equals the probability a true relationship exists, π , times the probability that the study has

sufficient power to obtain a significant estimate, $(1 - \beta)$. P_3 equals the probability a true relationship does not exist, $(1 - \pi)$, times the probability that Type I error produces a significant estimate, α .

(TABLE 1 HERE)

Now consider the set of all studies that estimate a statistically significant effect. This set consists of two types of studies: (i) those that have correctly discovered a true effect, and (ii) those that estimated a significant relationship when no relationship actually existed. The unconditional probabilities of each of these events occurring are $P_1 = \pi(1 - \beta)$ and $P_3 = (1 - \pi)\alpha$, respectively. Thus, the conditional probability that a real relationship exists given that a study reports a significant estimate -- what is called the “post-study probability that a relationship exists” -- is given by:

$$(1) \quad PSP(Relationship \text{ Exists}) = \frac{\pi(1-\beta)}{\pi(1-\beta) + (1-\pi)\alpha}.$$

While π and *Power* are inherently unobservable, researchers have put forward numbers that they think represent plausible values for these parameters. With respect to *Power*, numbers between 0.20 and 0.80 have been suggested as generally representative of studies in the social sciences.² Of course, nobody knows how many true relationships are “out there.” This is, after all, what studies are attempting to discover. Further, this is complicated in disciplines like economics where general equilibrium suggests that outcomes depend on a great many variables, at least to some extent.

However, if we restrict ourselves to considering relationships that are “economically significant”, where one variable has a quantitatively important effect on another, then a range of plausible values for π would arguably run from 0.01 (there is a 1 in a 100 chance that the

² Christensen and Miguel (2017, page 5) characterize 50% power as a “realistic level of statistical power for many studies.” Maniadis, Tufano, and List (2017) use 20% and 70% power in their numerical examples. Ioannidis, Doucouliagos, and Stanley (2017) estimate that the median statistical power in economics is approximately 18%.

studied relationship is real and substantive), to 0.50 (there is a 50% chance the relationship is real and substantive).³ Journals are generally not favorably disposed to publish studies where π is larger than 0.50 because the scientific value of these studies would be considered small (“everybody already knows this”).⁴

The top panel of Table 2 reports *PSP(Relationship Exists)* values for different π and *Power* values. For example, if we assume that the probability that the relationship being studied really exists is 0.10, and if the *Power* of the study is 0.50, then there is a 53% probability that the significant effect reported by the study is actually picking up a real effect.⁵

(TABLE 2 HERE)

Correspondingly, there is a 47% chance that the estimated, significant relationship is picking up something that is actually not there. This might seem surprising because the researcher set the significance level, α , at 0.05, so that we might expect to obtain a “false positive” only 5% of the time. But when 90 out of every 100 studies are looking for a relationship where none actually exists, the number of “false positives” will be disproportionately represented in the published literature. This is clearly seen when the probability of true relationships existing is even smaller, say 0.01. Now, 99 studies out of every 100 are looking for a relationship where none actually exists, producing a large number of “false positives” in the literature. In this case, only 9% of statistically significant relationships will be picking up a true effect. The other 91% are all reporting something that does not actually exist.

³ Benjamin et al. (2017, page 6) suggest an odds ratio of 1:10 for new discoveries in psychology and cancer clinical trials, with much lower odds for preclinical biomedical research.

⁴ Here we focus on studies whose contribution lies in identifying statistically significant relationships, as opposed to obtaining a precise estimate of a relationship that is already known to exist.

⁵ The results in Tables 2, 4, 5, and 7 can be replicated using a spreadsheet posted at Dataverse here: <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FMANV6Y>.

One aspect not explored in Ioannidis’ article is the *PSP* associated with an insignificant finding – what is sometimes called a “null” or “negative” result. The conditional probability that no relationship exists given that a study reports an insignificant estimate is given by:

$$(2) \quad PSP(\text{No Relationship Exists}) = \frac{(1-\pi)(1-\alpha)}{\pi\beta + (1-\pi)(1-\alpha)}.$$

The bottom panel of Table 2 reports *PSP(No Relationship Exists)* values for the same π and *Power* values reported in the top panel. Returning to the ($\pi = 0.10, \text{Power} = 0.50$) example from above, there is a 94% probability that an insignificant finding represents a true “no effect”. Note that over a relatively large range of π and *power* values, *PSP(No Relationship Exists) > PSP(Relationship Exists)*. In words, the probability that an insignificant finding in a published study indicates there really is no relationship is greater than the probability that a significant finding indicates that there is a relationship. Or to state it differently, an insignificant estimate in a published study is more “believable” than a significant one.

Students of STAT101 courses should find this last result counter-intuitive. It is often emphasized that failure to reject $H_0: \varphi_1 = 0$ should not be interpreted to mean $\varphi_1 = 0$ (“one should never accept the null hypothesis”), so that an insignificant estimate should not be interpreted to mean that there is no effect. On the other hand, rejection of the null allows one to accept the alternative hypothesis that $\varphi_1 \neq 0$, so that a significant estimate can be interpreted to mean that there is an effect. The results from TABLE 2 seem to turn this wisdom on its head.

The explanation for this apparent contradiction is that STAT101 and TABLE 2 are referring to different activities. The statistics of STAT101 is geared to interpreting the results of an individual experiment. In contrast, TABLE 2 is concerned with interpreting results that one sees in the published literature. As Ioannidis (2005) makes clear, the failure to appreciate

this distinction causes researchers to grossly misinterpret the results reported in empirical studies.

To this point, nothing has been said about “publication bias.” One might think, based on the results above, that journals would be most interested in publishing insignificant results, as these are, in some sense, more “reliable.” However, that is not the case. It is well known that journals prefer to report “important” and “novel” findings. This is often translated in practice to mean estimates that are statistically significant. Given this preference by journals, researchers, who are rewarded for publishing in journals, are motivated to produce statistically significant findings.

Accordingly, Ioannidis proceeds by introducing a fourth component into his analysis, *Bias*. *Bias* captures the effects of journal policies and researcher behaviors on the probability that a published research finding will be statistically significant. For example, journals may choose not to publish studies that have insignificant results because they are not considered scientifically “newsworthy.” This has the effect of filtering out insignificant results from the published literature, biasing downward the probability that published research findings are insignificant, and thus biasing upward the share of research findings that are significant.

These policies also affect researcher behavior. After obtaining an insignificant finding, some researchers may choose to give up on the research project, electing not to write up the results and submit them to a journal, knowing that their research is unlikely to be published. This is often referred to as the “file drawer” effect. Alternatively, researchers can work the data more intensively to try and produce a significant estimate. The procedures by which this is done are referred to by colorful terms such as “data mining”, “p-hacking”, and “the garden of the forking paths”. If one does not find a significant effect, one can try alternative approaches such as substituting other variables in the equation, eliminating observations that are viewed as “unusual” (“outliers”), or experimenting with alternative estimation procedures. Researchers

keep going until they obtain a significant estimate, and it is that estimate which gets published. All of these practices at the journal and individual researcher level are combined in the concept of *Bias*.

Let μ represent the decreased share of insignificant estimates that appear in the published literature due to *Bias*, where the reference point is a world with no *Bias*. A simple adjustment to the TABLE 1 probabilities allows one to determine how *Bias* alters the resulting *PSP* values. For example, in the absence of *Bias*, the probability that a published study reports an insignificant relationship when a relationship truly exists is $\pi\beta$. With *Bias*, this probability falls to $\pi\beta(1 - \mu)$. Concurrently, the probability that a published study reports a significant finding when a relationship really exists is $\pi(1 - \beta) + \pi\beta\mu$. A similar calculation adjusts the probabilities when a relationship does not exist.

(TABLE 3 HERE)

The corresponding post-study probabilities in the presence of *Bias* are given by:

$$(3) \quad PSP(Relationship \text{ Exists}/Bias) = \frac{\pi(1-\beta) + \pi\beta\mu}{\pi(1-\beta) + \pi\beta\mu + (1-\pi)\alpha + (1-\pi)(1-\alpha)\mu}.$$

and

$$(4) \quad PSP(No \text{ Relationship Exists}/Bias) = \frac{(1-\pi)(1-\alpha)(1-\mu)}{\pi\beta(1-\mu) + (1-\pi)(1-\alpha)(1-\mu)}.$$

TABLE 4 repeats the analysis of TABLE 2 for a variety of *Bias* values, focusing on *PSP(Relationship Exists)*. The top panel reproduces the no *Bias* case ($\mu = 0$) from TABLE 2 to facilitate comparison. The next three panels report *PSP(Relationship Exists)* values for increasing degrees of *Bias*: $\mu = 0.10$, $\mu = 0.25$, and $\mu = 0.50$.⁶

(TABLE 4 HERE)

⁶ Maniadis et al. (2017) assume a *Bias* value of 0.40 in their numerical calculations for “adversarial bias.”

Even a relatively small degree of *Bias* can have a substantial effect. For example, compare the difference in *PSP* values when $Bias = 0.0$ and $Bias = 0.10$ for the case when ($\pi = 0.10, Power = 0.50$). This relatively small increase in *Bias* reduces $PSP(Relationship\ Exists)$ from 0.53 to 0.30. In words, the probability that a relationship exists given that a study reports a significant finding falls from approximately half to less than a third when journal and researcher bias reduce the share of insignificant findings by 10%.

Many researchers would argue that *Bias* is likely to be greater than 0.10 in the real world of academic publishing.⁷ The subsequent panels consider increasing values of *Bias*. Continuing with the case ($\pi = 0.10, Power = 0.50$), $PSP(Relationship\ Exists)$ falls from 0.53 to 0.19 and 0.14 as *Bias* increases from $\mu = 0$ to $\mu = 0.25$ and $\mu = 0.50$, respectively. Given a 14% probability that a significant finding indicates that a relationship actually exists, it is not hard to understand why Ioannidis entitled his paper, “Why Most Published Research Findings Are False.”

III. Some Proposed Fixes to the Reproducibility Crisis That Do Not Involve Replication

A number of suggestions have been made to address the “reproducibility” crisis. Most of these can be fit within the framework above, with the majority directed at trying to reduce *Bias*.

Publish insignificant findings. The most straightforward approach is for journals to be willing to publish “research failures”, i.e., null or negative results. This would directly decrease *Bias* by allowing more insignificant estimates into the literature. This, in turn, would diminish the incentive for researchers to data mine for significant results. There are regular calls for journals to do this (Menclova, 2017). However, previous efforts to start journals dedicated to

⁷ In a recent study, Andrews and Kasy (2017) estimate that significant findings at the 5% level are 30 times more likely to be published than insignificant findings.

publishing negative results have not been very successful.⁸ To date, there are no journals in economics that are dedicated to negative results.⁹

Pre-Registration. Pre-registration is a public declaration of intention where the researcher states what they intend to study and the hypotheses they plan to investigate. Registration is made before data are collected and analyzed. Pre-registration is designed to address the “file drawer” problem -- that studies are begun but never completed because the results did not turn out sufficiently “favorably”. Faced with insignificant results that are unlikely to get published, researchers may not invest the additional work to write up the results and submit them to a journal. Franco, Malhotram and Simonovits (2014) present evidence that this, in fact, is the main source of “publication bias.” Pre-registration does not force the researcher to follow through on their study through publication, but it hopefully creates a greater incentive to do so. Further, it lets other researchers know that a project was begun but not completed, and that can be useful information in and of itself.

Pre-registration is standard procedure in medical randomized controlled trials (RCTs) and is becoming more common in the social sciences and economics. The American Economic Association (AEA) has a registry for posting about RCTs.¹⁰ Other organizations supporting pre-registration registries are Evidence in Governance and Politics (EGAP)¹¹, International Initiative for Impact Evaluation (3ie)¹², and the Open Science Framework.¹³

Pre-Analysis Plans. Pre-Analysis Plans (PAPs) are a subset of pre-registrations but are distinguished because they offer greater detail about the researcher’s plans. Data collection is

⁸ In recent years, a number of journals were founded that were dedicated to publishing negative results, such as *New Negatives in Plant Science*, *Journal of Negative Results*, *Journal of Negative Results in BioMedicine*, *Journal of Pharmaceutical Negative Results*, and *The All Results Journals*. However, most of these journals have either stopped publishing, or publish infrequently, as can be attested by clicking on the links in this online article: Goodchild van Hilten (2015).

⁹ The journal *FinanzArchiv/Public Finance Analysis* recently put out a call for a special issue on “Insignificant Results in Public Finance” (Weichenrieder, 2017).

¹⁰ See <http://socialscienceregistry.org>.

¹¹ See <http://egap.org/design-registration>.

¹² See <http://ridie.3ieimpact.org>.

¹³ See <http://osf.io>.

more thoroughly described. The exact hypotheses the researcher plans to test are specified in advance. And rules about how data will be handled (e.g., elimination of “outliers”) are spelled out before actual data analysis. While registration is designed to bring into the light studies that might otherwise remain unseen, PAPs are designed to directly affect *Bias*. Specifically, they are designed to tie the researcher’s hands before coming to the data. The result should be less data mining and p-hacking, which should reduce μ .

Registered Reports. Registered reports go further than PAPs because they are designed to tie both the researcher’s and the journal’s hands. In a registered report, a researcher submits a detailed study plan to a journal before undertaking data collection and analysis. The journal puts the plan out to review and reviewers decide whether to accept the study in principle before the analysis is carried out. The reviewers and the journal base their decision on the importance of the question and the appropriateness and promise of the researcher’s plan of study to be able to answer the research question. At this stage, reviewers can still influence the study by suggesting improvements in the researcher’s study plan. After the research is carried out, reviewers again assess the study for journal publication, but their decision should be based solely on whether the researcher faithfully executed their study plan. The decision is supposed to be independent of the actual results. Thus, registered reports focus on the inputs to the research process rather than the outputs.

The use of registered reports is growing impressively. There are currently over 80 journals that either institute registered reports as part of their normal submission process, or have sponsored special issues in which all the studies followed the registered reports paradigm. A list of the journals that support registered reports can be found at <https://cos.io/rr/>. At the date of this writing, there are no economics journals on the list.

Results-blind or Results-free Review. Results-free or results-blind reviewing is very similar to registered reports except all the research work is completed at the time of submission.

The researcher sends in the completed paper, minus the results and conclusions. In principle, reviewers are supposed to give an up-or-down assessment of the study without knowing how the results turned out. If the journal decides to accept the manuscript, the researcher adds the results and conclusions. The manuscript then undergoes a final review, but only to ensure that the interpretation of the results is appropriate.

Like PAPs, registered reports and results-blind review are designed to directly diminish the size of μ . They do this in two ways. First, they try to limit the degree of data mining and analysis manipulation on the part of the researcher. In the case of registered reports, they do this by requiring the researcher to specify their study design in advance, thus restricting the potential for future manipulation. In the case of results-blind reviewing, they do this by taking away the incentive to data mine and p -hack. Since the journal will make their decision without seeing the results, there is no need for the researcher to try and manipulate the data to get a certain outcome. And in both cases, journals make their acceptance decision without knowing the outcome of the research.

Lowering the significance level. An alternative approach that has recently received much attention is a call to lower the level of significance from 0.05 to 0.005 (Benjamin et al., 2017). While there are different motivations for doing this, most researchers believe that this will make it easier to identify “real relationships.”

TABLE 5 shows how this would work. The top two panels repeat the preceding analysis for the case where there is no bias ($Bias = 0$), first with a significance level of $\alpha = 0.05$, then with the stricter standard of $\alpha = 0.005$. Because this will reduce the likelihood of obtaining a false positive, one would expect that those estimates that are significant at this standard will be more likely to represent real effects. In fact, that is exactly what one sees in these top two panels. If we again go to our reference case of ($\pi = 0.10, Power = 0.50$), we see that an α of

0.05 results in a $PSP(Relationship\ Exists)$ value of 0.53. With an α value of 0.005, this rises to 0.92. One sees similar increases for many other $(\pi, Power)$ combinations.

(TABLE 5 HERE)

However, as the next two panels of TABLE 5 illustrate, the existence of *Bias* can greatly diminish the improvements from instituting a stricter level of significance. In the case of $(\pi = 0.10, Power = 0.50)$, decreasing the level of significance from 0.05 to 0.005 results in only a very small improvement in $PSP(Relationship\ Exists)$, from 0.19 to 0.21. While the lower value of α produces a lot more insignificant estimates. *Bias* squelches these insignificant estimates and increases the share of (false) significant ones. Thus, a sufficiently high degree of *Bias* can undermine much of the good that comes from lowering α .¹⁴

Advocates of lowering α might counter that decreasing α would also have the effect of decreasing *Bias*, since it would make it harder to *p*-hack one's way to a significant result if no relationship really exists. However, lowering α would also lower *Power*, since it will be harder even for true relationships to achieve significance. Just how all these consequences of lowering α would play out in practice is not known, but the last two panels of TABLE 5 present a less than sanguine picture. Suppose that before the change in α , $Bias = 0.25$ and $Power = 0.50$. Lowering α from 0.05 to 0.005 decreases *Bias* and *Power*. Suppose that the new values are $Bias = 0.15$ and $Power = 0.20$. A comparison of these two panels shows that the ultimate effect of decreasing α on PSP is approximately zero.

IV. How Replication Affects the Post-Study Probability that a Relationship Exists

As we shall discuss below, there are many different types of replications. Very generally, a replication involves repeating the analysis of an original study to confirm its trustworthiness. We can use the framework above to show how replications can substantially improve our

¹⁴ In their call to lower α to 0.005, Benjamin et al. (2017) do not consider the role that *Bias* plays in affecting post-study probabilities.

understanding of whether a true relationship exists. As above, let PSP represent the probability that a true relationship exists given that a study reports a significant finding. Suppose now a replication of that study is done, using the exact same procedures as the original study, and drawing observations from the same population. We assume for the time being that the replication process is devoid of bias.

Consider first the outcome that a replication successfully replicates the original study, defined as obtaining a significant estimate at the same level of significance ($\alpha = 0.05$). There are two ways this outcome can occur. First, a real relationship exists and the replication produces a significant estimate. The probability of this event occurring is $PSP \cdot (1 - \beta)$, where PSP is the post-study probability that a true relationship exists, and $(1 - \beta)$ is the power of the replication. The other possibility is that a true relationship does not exist, but Type I error yields a significant result. The associated probability is $(1 - PSP) \cdot \alpha$. Accordingly, the post-replication probability that a true relationship exists, conditional on the replication being a success, is given by

$$(5) \quad PSP(\text{Replication Successful}/\text{No Bias}) = \frac{PSP \cdot (1 - \beta)}{PSP \cdot (1 - \beta) + (1 - PSP) \cdot \alpha}.$$

Now consider the outcome that a replication does not successfully replicate an original study. This can occur because a relationship really does exist but the replication is not sufficiently powerful to identify it. This probability is given by $PSP \cdot \beta$. Alternatively, the null hypothesis of no relationship may fail to be rejected because it is true. This probability is given by $(1 - PSP)(1 - \alpha)$. Accordingly, the post-replication probability that a true relationship exists after failing to obtain a successful replication result is given by

$$(6) \quad PSP(\text{Replication Unsuccessful}/\text{No Bias}) = \frac{PSP \cdot \beta}{PSP \cdot \beta + (1 - PSP)(1 - \alpha)}.$$

As before, our analysis can accommodate the influence of *Bias*. Maniadis, Tufano, and List (2017) discuss several types of bias that can arise in a replication. Here we consider the

case of “adversarial bias”, in which replication researchers are biased towards showing that an original study is wrong. Besides pure misanthropy, replication researchers may be influenced by the fact that it is well-known that journals are more likely to publish replications that dispute, rather than confirm, the findings of an original study.¹⁵ Let this *Adversarial Bias* be represented by the parameter ω . The associated probabilities are reported in TABLE 6.

(TABLE 6 HERE)

From there, it follows that the post-replication probabilities with *Adversarial Bias* are given by:

$$(7) \quad PSP(Replication\ Successful|Adversarial\ Bias) = \frac{PSP \cdot (1-\beta)(1-\omega)}{PSP \cdot (1-\beta)(1-\omega) + (1-PSP) \cdot \alpha(1-\omega)}.$$

and

$$(8) \quad PSP(Replication\ Unsuccessful|Adversarial\ Bias) = \frac{PSP \cdot \beta + PSP \cdot (1-\beta)\omega}{PSP \cdot \beta + PSP \cdot (1-\beta)\omega + (1-PSP)(1-\alpha) + (1-PSP) \cdot \alpha\omega}.$$

TABLE 7 reports the respective post-replication probabilities for a variety of *PSP* and *Power* values. The *PSP* values identify the probabilities that a relationship exists after an original study reports a significant finding, but before a replication is done. The values in the table update this probability depending on whether the replication was successful (*Replication Success* = 1) or unsuccessful (*Replication Success* = 0).

(TABLE 7 HERE)

Consider the following scenario: A researcher studies a phenomenon that has a 0.10, pre-study, *ex ante* probability of being true ($\pi = 0.10$). The study has *Power* = 0.50, and publication and researcher bias favor the reporting of significant results, so that $\mu = 0.25$. Given these parameters, the post-study probability that a relationship exists after an original

¹⁵ While the *American Economic Review* has published many replications, it has never published a replication that confirms the original study (Duvendack, Palmer-Jones, and Reed, 2015).

study reports a significant result (PSP) is only 19% (see TABLE 5). Along comes another researcher who does a replication study of the same phenomenon. Let us assume for the moment that this researcher is unbiased. Let's also round up the PSP to 0.20 and assume that the replication study has the same power as the original study ($Power = 0.50$).

The values in the table show the original PSP of 0.20 gets updated depending on whether the replication was unsuccessful or successful. Following an unsuccessful replication, the post-replication probability that a relationship exists falls from 0.20 to 0.12. However, a successful replication raises the probability from 0.20 to 0.71.

This result is not unusual. As demonstrated in TABLE 7, the difference in post-replication probabilities for successful and unsuccessful replications is substantial across a wide variety of PSP and $Power$ values. Further, introducing *Adversarial Bias* has little effect on the results. When an *Adversarial Bias* of 0.25 is introduced to the ($PSP=0.20$, $Power=0.50$) case, the corresponding post-replication probabilities for an unsuccessful and successful replication are 0.14 and 0.71, respectively. When *Adversarial Bias* increases to 0.50 (not shown), the corresponding values are 0.16 and 0.71. This analysis highlights the important role that replication can play in addressing the reproducibility crisis in science.

V. Replications in Greater Detail

There is no black and white line separating replications from original studies. A study reports that a treatment x increases y . Another study is also interested in the effect of x on y . How dissimilar from the first study does the second study have to be before it ceases to be a replication? The answer is unclear.

Nor is there a clear criterion of “success” when it comes to replications. Suppose an original study estimates that a treatment x increases y by 5%, with a confidence interval of [3%,7%]. A replication study uses the exact same procedures as the first study, with data drawn from the same population, and also estimates a treatment effect of 5%, but the estimate is

insignificant with a confidence interval of [-1%,11%]. Has the replication study replicated the original study, or not? Suppose the replication study estimated a larger effect, though still insignificant. Was the replication a success? Or a failure?

It is because of these ambiguities that Duvendack, Palmer-Jones, and Reed (2015) define a replication as any study whose primary purpose is to determine whether a previously published study is true. They define “replication success” as any result that would have led the replication authors to make the same conclusion as the original study, and “replication failure” as any result that contradicts the conclusion of the original study.

Just as there is no standard definition of a replication, nor generally accepted criteria of “replication success,” nor is there any consensus on how best to categorize the many different types of replications. Reed (2017) identifies 6 types of replications.

Type 1: Reproduction. As the name suggests, *Reproduction* attempts to reproduce the results from an original study using the same data and methods. This type of replication is done primarily to detect whether the results reported in the original analysis contain any errors.

Type 2: Robustness Analysis – Same Dataset. *Robustness Analysis – Same Dataset* is similar to *Reproduction* in that it uses the exact same data. However, it goes beyond the original study by making modifications to the data or the methods of analysis in order to see if the results are robust to reasonable changes to the original analysis. This type of replication is done primarily to determine if other researchers, given the same data as the original study, would likely have come to the same conclusion.

Type 3: Repetition. *Repetition* is also similar to *Reproduction* in that it uses the same methods of analysis as the original study. However, while it targets the same population, it uses a different sample. A good example of a repetition would be a researcher who replicates a famous economic experiment using a different sample of subjects who share the same characteristics as the original study (age, gender, occupation, etc.) This type of study is

primarily done to investigate whether sampling error could be responsible for the results of the original study.

Type 4: Extension. *Extension* is similar to *Repetition* in that it uses the same methods of analysis. It deviates in that it draws its data from a different, though usually related, population. For example, a study might find that white males respond to a certain treatment. The replication might use the same procedures to see if white females respond the same way to the treatment. This type of replication is primarily done to determine the boundaries of external validity for the original study.

Types 5 & 6: Robustness Analysis – Same Population & Robustness Analysis – Different Population. *Robustness Analysis – Same Population & Robustness Analysis – Different Population* explore modifications to the data in *Repetition* and *Extension*, respectively, to see how robust the results are to reasonable modifications to the data and methods of analysis. Like *Robustness Analysis – Same Dataset*, these types of replications are designed primarily to determine if other researchers, were they given the data from a *Repetition* or *Extension* replication, would likely come to the same conclusion.

(FIGURE 1 HERE)

In practice, it is very common to see multiple types of replications included in a single replication study. For example, a replication might first reproduce the original findings using the same data and code as the original study, and then check different variable specifications for robustness (Types 1 and 2). It might then extend the dataset to different countries or time periods (Type 5).

A major distinction can be made between replications that use the same data as the original study, and those that draw their data from other sources. The former case is sometimes called an “internal replication” (Wood and Vasquez, 2017). Internal replications limit the options available to the replicating researcher. For example, a researcher may want to explore

the issue of endogeneity, but no viable instruments are included in the original data. Opening up the data creates more opportunities for robustness checking, but this also expands the “garden of forking paths.”

This highlights the somewhat artificial nature of the post-replication probabilities analysis above. For example, power is an irrelevant concept in a Type 1 replication study, since this is essentially a checking exercise to make sure that numbers are correctly reported. Fundamental to the probabilities derived in Equations (7) and (8) is the element of statistical uncertainty. In this sense, the *PSP* values reported in TABLE 7 are most appropriate for Type 3 replications, where identical procedures are applied to data drawn from the same population as the original study. The further replications deviate from the Type 3 model, the less applicable are the associated probabilities. Even so, the values in TABLE 7 are illustrative of the potential for replication to substantially alter the probability that a significant estimate represents a true relationship.

VI. Conclusion

Reproducibility is at the heart of the scientific enterprise. Presumably, what defines science is its ability to discover relationships in nature and behavior. For a very long time, it was generally assumed that once a study had reported a significant finding, that finding could be viewed as reliable, as representing a real phenomenon. Ioannidis’ (2005) seminal study made it clear that there is a distinction between the probability that a finding has uncovered a real relationship at the level of an individual experiment, and the probability that a significant finding reported in the academic literature represents a real relationship.

This article uses Ioannidis’ framework to explain why many researchers believe that there is a “reproducibility crisis” in science. It then goes on to use that framework to evaluate various proposals to fix the problem. Of particular interest in our study is the “post-study probability” (*PSP*), the probability that a reported research finding represents a true

relationship. This probability is inherently unknowable, as it depends on a number of parameters that cannot be observed. However, several insightful results emerge if we are willing to make some conjectures about reasonable parameter values.

First, under reasonable parameter values, a published finding that is statistically insignificant is generally more likely to indicate that there really is no effect, than a statistically significant finding is to indicate that there is an effect. Second, publication bias, defined as the preference of both journals and researchers to report significant results, exercises a large influence on *PSP*. Even small amounts of bias can greatly diminish *PSPs*, causing the published literature to be unreliable. Most solutions to the reproducibility crisis focus on ways to reduce publication bias. An interesting variant of these is the recent call to lower the significance level from 0.05 to 0.005 (Benjamin et al., 2017). In the presence of publication bias, lowering the significance level, by itself, is unlikely to have much effect on *PSPs*. It will need to substantially reduce publication bias in order to be effective. Finally, a single replication, if it confirms the original study, can have a large impact on the probability that the estimated effect is real. To be fair, there are many types of replications, and this result strictly applies to only one of them. However, it does provide some optimism for believing that replication can play an important role in ameliorating the reproducibility crisis.

References

American Economic Review, (2017). Vol. 107, No. 5.

Andrews, I. and Kasy, M. (2017) Identification and correction for publication bias. Working paper 23298, National Bureau of Economic Research, November 2017.
Retrieved from <http://www.nber.org/papers/w23298>.

Benjamin, D.J., Berger, J.O., Johannesson, M. Nosek, B.A., Wagenmakers, E.-J., Berk, R., ..., Johnson, V.E. (2017). Redefine statistical significance. *Nature Human Behaviour*, 1(0189).
Retrieved from <https://www.nature.com/articles/s41562-017-0189-z>.

Christensen, G.S. and Miguel, E. (2016). Transparency, reproducibility, and the credibility of economics research. CEGA Working Paper Series No. WPS-065. Center for Effective Global Action. University of California, Berkeley.

Duvendack, M., Palmer-Jones, R.W. and Reed, W.R. (2015). Replications in economics: A progress report. *Econ Journal Watch*, 12(2): 164-191.

Duvendack, M., Palmer-Jones, R.W. and Reed, W.R. (2017). What is meant by ‘replication’ and why does it encounter resistance in economics? *American Economic Review*, 107(5): 46-51.

Franco, A., Malhotra, N., and Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science* 345(6203): 1502-5.

Goodchild van Hilten, L. (2015, May 5). Publishing bias favors positive results; now there’s a movement to change that. [Blog post] Retrieved from www.elsevier.com/connect/scientists-we-want-your-negative-results-too.

Ioannidis, J.P. (2005). Why most published research findings are false. *PloS Medicine*, 2(8): 1418-1422.

Ioannidis, J.P., Doucouliagos, H. and Stanley, T. (2017). The power of bias in economics. *Economic Journal* 127(605): F236-65.

Maniadis, Z., Tufano, F., and List, J.A. (2017). To replicate or not to replicate? Exploring reproducibility in economics through the lens of a model and a pilot study. *Economic Journal*, 127(605): F209-F235.

Menclova, A. (2017, March 24). Is it time for a journal of insignificant results? [Blog post]. Retrieved from <https://replicationnetwork.com/2017/03/24/menclova-is-it-time-for-a-journal-of-insignificant-results/>.

Reed, W.R. (2017). Replication in labor economics. *IZA World of Labor*, 413. doi: 10.15185/izawol.413

Wood, B.D.K., and Vasquez, M. (2017). Microplots and food security: Encouraging replication studies of policy relevant research. Economics Discussion Papers, No 2017-71, Kiel Institute for the World Economy.
<http://www.economics-ejournal.org/economics/discussionpapers/2017-71>

TABLE 1
Joint Probabilities of a Relationship Existing and Obtaining a Significant Estimate

<i>Statistically Significant Estimate?</i>	<i>True Relationship Exists?</i>	
	<i>Yes</i>	<i>No</i>
<i>Yes</i>	$P_1 = \pi(1 - \beta)$	$P_3 = (1 - \pi)\alpha$
<i>No</i>	$P_2 = \pi\beta$	$P_4 = (1 - \pi)(1 - \alpha)$

NOTE: π is the probability that the relationship being studied actually exists. α is the Type I error associated with the null hypothesis that a relationship does not exist. β is the Type II error associated with the null hypothesis that a relationship does not exist, so that $(1 - \beta)$ is the associated power of the study.

TABLE 2
Post-Study Probabilities (PSPs) for Different π and Power Values

π	<i>Power</i>				
	<i>0.20</i>	<i>0.30</i>	<i>0.50</i>	<i>0.70</i>	<i>0.80</i>
<i>Post-Study Probability (Relationship Exists)</i>					
<i>0.01</i>	0.04	0.06	0.09	0.12	0.14
<i>0.05</i>	0.17	0.24	0.34	0.42	0.46
<i>0.10</i>	0.31	0.40	0.53	0.61	0.64
<i>0.20</i>	0.50	0.60	0.71	0.78	0.80
<i>0.30</i>	0.63	0.72	0.81	0.86	0.87
<i>0.40</i>	0.73	0.80	0.87	0.90	0.91
<i>0.50</i>	0.80	0.86	0.91	0.93	0.94
<i>Post-Study Probability (No Relationship Exists)</i>					
<i>0.01</i>	0.99	0.99	0.99	1.00	1.00
<i>0.05</i>	0.96	0.96	0.97	0.98	0.99
<i>0.10</i>	0.91	0.92	0.94	0.97	0.98
<i>0.20</i>	0.83	0.84	0.88	0.93	0.95
<i>0.30</i>	0.73	0.76	0.82	0.88	0.92
<i>0.40</i>	0.64	0.67	0.74	0.83	0.88
<i>0.50</i>	0.54	0.58	0.66	0.76	0.83

NOTE: π is the probability that the relationship being studied actually exists. *Power* is the probability that a study will produce a statistically significant estimate given that a relationship actually exists. The values for *Post-Study Probability (Relationship Exists)* and *Post-Study Probability (No Relationship Exists)* are calculated using Equations (1) and (2) in the text, respectively. Values can be checked, and alternative parameter values investigated, using the spreadsheet posted at:

<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FMANV6Y>.

TABLE 3
Joint Probabilities of a Relationship Existing and Obtaining a Significant Estimate
in the Presence of Bias

<i>Statistically Significant Estimate?</i>	<i>True Relationship Exists?</i>	
	<i>Yes</i>	<i>No</i>
<i>Yes</i>	$P_1 = \pi(1 - \beta) + \pi\beta\mu$	$P_3 = (1 - \pi)\alpha + (1 - \pi)(1 - \alpha)\mu$
<i>No</i>	$P_2 = \pi\beta(1 - \mu)$	$P_4 = (1 - \pi)(1 - \alpha)(1 - \mu)$

NOTE: π is the probability that the relationship being studied actually exists. α is the Type I error associated with the null hypothesis that a relationship does not exist. β is the Type II error associated with the null hypothesis that a relationship does not exist, so that $(1 - \beta)$ is the associated power of the study. μ represents the decreased share of insignificant estimates that appear in the published literature due to publication bias.

TABLE 4
PSP(Relationship Exists) for Different π , Power, and Bias Values

π	<i>Power</i>				
	<i>0.20</i>	<i>0.30</i>	<i>0.50</i>	<i>0.70</i>	<i>0.80</i>
<i>Bias (μ) = 0.00</i>					
<i>0.01</i>	0.04	0.06	0.09	0.12	0.14
<i>0.05</i>	0.17	0.24	0.34	0.42	0.46
<i>0.10</i>	0.31	0.40	0.53	0.61	0.64
<i>0.20</i>	0.50	0.60	0.71	0.78	0.80
<i>0.30</i>	0.63	0.72	0.81	0.86	0.87
<i>0.40</i>	0.73	0.80	0.87	0.90	0.91
<i>0.50</i>	0.80	0.86	0.91	0.93	0.94
<i>Bias (μ) = 0.10</i>					
<i>0.01</i>	0.02	0.03	0.04	0.05	0.05
<i>0.05</i>	0.09	0.12	0.17	0.21	0.23
<i>0.10</i>	0.18	0.22	0.30	0.36	0.39
<i>0.20</i>	0.33	0.39	0.49	0.56	0.59
<i>0.30</i>	0.45	0.52	0.62	0.68	0.71
<i>0.40</i>	0.56	0.63	0.72	0.77	0.79
<i>0.50</i>	0.66	0.72	0.79	0.83	0.85
<i>Bias (μ) = 0.25</i>					
<i>0.01</i>	0.01	0.02	0.02	0.03	0.03
<i>0.05</i>	0.07	0.08	0.10	0.12	0.13
<i>0.10</i>	0.13	0.16	0.19	0.23	0.25
<i>0.20</i>	0.26	0.29	0.35	0.40	0.43
<i>0.30</i>	0.37	0.41	0.48	0.54	0.56
<i>0.40</i>	0.48	0.52	0.59	0.64	0.66
<i>0.50</i>	0.58	0.62	0.68	0.73	0.75
<i>Bias (μ) = 0.50</i>					
<i>0.01</i>	0.01	0.01	0.01	0.02	0.02
<i>0.05</i>	0.06	0.06	0.07	0.08	0.08
<i>0.10</i>	0.11	0.12	0.14	0.15	0.16
<i>0.20</i>	0.22	0.24	0.26	0.29	0.30
<i>0.30</i>	0.33	0.35	0.38	0.41	0.42
<i>0.40</i>	0.43	0.45	0.49	0.52	0.53
<i>0.50</i>	0.53	0.55	0.59	0.62	0.63

NOTE: π is the probability that the relationship being studied actually exists. *Power* is the probability that a study will produce a statistically significant estimate given that a relationship actually exists. μ represents the decreased share of insignificant estimates that appear in the published literature due to publication bias. The values for *Post-Study Probability (Relationship Exists)* are calculated using Equation (3) in the text. Values can be checked, and alternative parameter values investigated, using the spreadsheet posted at: <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FMANV6Y>.

TABLE 5
***PSP(Relationship Exists)* for Different π , Power, and Bias Values:**
The Effect of Lowering α to 0.005

π	<i>Power</i>				
	<i>0.20</i>	<i>0.30</i>	<i>0.50</i>	<i>0.70</i>	<i>0.80</i>
<i>Bias = 0, $\alpha = 0.05$</i>					
<i>0.01</i>	0.04	0.06	0.09	0.12	0.14
<i>0.05</i>	0.17	0.24	0.34	0.42	0.46
<i>0.10</i>	0.31	0.40	0.53	0.61	0.64
<i>0.20</i>	0.50	0.60	0.71	0.78	0.80
<i>0.30</i>	0.63	0.72	0.81	0.86	0.87
<i>0.40</i>	0.73	0.80	0.87	0.90	0.91
<i>0.50</i>	0.80	0.86	0.91	0.93	0.94
<i>Bias = 0, $\alpha = 0.005$</i>					
<i>0.01</i>	0.29	0.38	0.50	0.59	0.62
<i>0.05</i>	0.68	0.76	0.84	0.88	0.89
<i>0.10</i>	0.82	0.87	0.92	0.94	0.95
<i>0.20</i>	0.91	0.94	0.96	0.97	0.98
<i>0.30</i>	0.94	0.96	0.98	0.98	0.99
<i>0.40</i>	0.96	0.98	0.99	0.99	0.99
<i>0.50</i>	0.98	0.98	0.99	0.99	0.99
<i>Bias = 0.25, $\alpha = 0.05$</i>					
<i>0.01</i>	0.01	0.02	0.02	0.03	0.03
<i>0.05</i>	0.07	0.08	0.10	0.12	0.13
<i>0.10</i>	0.13	0.16	0.19	0.23	0.25
<i>0.20</i>	0.26	0.29	0.35	0.40	0.43
<i>0.30</i>	0.37	0.41	0.48	0.54	0.56
<i>0.40</i>	0.48	0.52	0.59	0.64	0.66
<i>0.50</i>	0.58	0.62	0.68	0.73	0.75
<i>Bias = 0.25, $\alpha = 0.005$</i>					
<i>0.01</i>	0.02	0.02	0.02	0.03	0.03
<i>0.05</i>	0.08	0.09	0.11	0.14	0.15
<i>0.10</i>	0.15	0.17	0.21	0.25	0.27
<i>0.20</i>	0.28	0.32	0.38	0.43	0.46
<i>0.30</i>	0.40	0.45	0.51	0.57	0.59
<i>0.40</i>	0.51	0.56	0.62	0.67	0.69
<i>0.50</i>	0.61	0.65	0.71	0.75	0.77

π	<i>Power</i>				
	<i>0.20</i>	<i>0.30</i>	<i>0.50</i>	<i>0.70</i>	<i>0.80</i>
<i>Bias = 0.25, $\alpha = 0.05$</i>					
<i>0.01</i>	0.01	0.02	0.02	0.03	0.03
<i>0.05</i>	0.07	0.08	0.10	0.12	0.13
<i>0.10</i>	0.13	0.16	0.19	0.23	0.25
<i>0.20</i>	0.26	0.29	0.35	0.40	0.43
<i>0.30</i>	0.37	0.41	0.48	0.54	0.56
<i>0.40</i>	0.48	0.52	0.59	0.64	0.66
<i>0.50</i>	0.58	0.62	0.68	0.73	0.75
<i>Bias = 0.15, $\alpha = 0.005$</i>					
<i>0.01</i>	0.02	0.03	0.04	0.05	0.05
<i>0.05</i>	0.10	0.12	0.16	0.20	0.22
<i>0.10</i>	0.19	0.23	0.29	0.35	0.37
<i>0.20</i>	0.34	0.40	0.48	0.55	0.57
<i>0.30</i>	0.47	0.53	0.62	0.67	0.70
<i>0.40</i>	0.58	0.64	0.71	0.76	0.78
<i>0.50</i>	0.67	0.72	0.79	0.83	0.84

NOTE: π is the probability that the relationship being studied actually exists. *Power* is the probability that a study will produce a statistically significant estimate given that a relationship actually exists. *Bias* represents the decreased share of insignificant estimates that appear in the published literature due to publication bias. The values in the table represent the *Post-Study Probability (Relationship Exists)* and are calculated using Equation (3) in the text. Values can be checked, and alternative parameter values investigated, using the spreadsheet posted at: <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FMANV6Y>.

TABLE 6
Joint Probabilities of a Relationship Existing and Obtaining a Successful Replication
in the Presence of Adversarial Bias

<i>Successful Replication?</i>	<i>True Relationship Exists?</i>	
	<i>Yes</i>	<i>No</i>
<i>Yes</i>	$P_1 = PSP \cdot (1 - \beta)(1 - \omega)$	$P_3 = (1 - PSP) \cdot \alpha(1 - \omega)$
<i>No</i>	$P_2 = PSP \cdot \beta + PSP \cdot (1 - \beta)\omega$	$P_4 = (1 - PSP)(1 - \alpha) + (1 - PSP) \cdot \alpha\omega$

NOTE: *PSP* is the probability that the relationship being studied actually exists following the results from an original, published study. α is the Type I error associated with the null hypothesis that a relationship does not exist. β is the Type II error associated with the null hypothesis that a relationship does not exist, so that $(1 - \beta)$ is the associated power of the replication study. ω represents the decreased share of significant estimates that appear among replication studies in the published literature due to adversarial bias.

TABLE 7
The Effect of a Replication on the Post-Study Probability That a Relationship Exists for
Different *PSP*, *Power*, and *Adversarial Bias* Values

<i>PSP</i>	<i>Replication Success</i>	<i>Power</i>				
		<i>0.20</i>	<i>0.30</i>	<i>0.50</i>	<i>0.70</i>	<i>0.80</i>
<i>Adversarial Bias = 0</i>						
<i>0.10</i>	<i>0</i>	0.09	0.08	0.06	0.03	0.02
	<i>1</i>	0.31	0.40	0.53	0.61	0.64
<i>0.20</i>	<i>0</i>	0.17	0.16	0.12	0.07	0.05
	<i>1</i>	0.50	0.60	0.71	0.78	0.80
<i>0.30</i>	<i>0</i>	0.27	0.24	0.18	0.12	0.08
	<i>1</i>	0.63	0.72	0.81	0.86	0.87
<i>0.40</i>	<i>0</i>	0.36	0.33	0.26	0.17	0.12
	<i>1</i>	0.73	0.80	0.87	0.90	0.91
<i>0.50</i>	<i>0</i>	0.46	0.42	0.34	0.24	0.17
	<i>1</i>	0.80	0.86	0.91	0.93	0.94
<i>0.60</i>	<i>0</i>	0.56	0.53	0.44	0.32	0.24
	<i>1</i>	0.86	0.90	0.94	0.95	0.96
<i>0.70</i>	<i>0</i>	0.66	0.63	0.55	0.42	0.33
	<i>1</i>	0.90	0.93	0.96	0.97	0.97
<i>0.80</i>	<i>0</i>	0.77	0.75	0.68	0.56	0.46
	<i>1</i>	0.94	0.96	0.98	0.98	0.98
<i>0.90</i>	<i>0</i>	0.88	0.87	0.83	0.74	0.65
	<i>1</i>	0.97	0.98	0.99	0.99	0.99
<i>Adversarial Bias = 0.25</i>						
<i>0.10</i>	<i>0</i>	0.09	0.08	0.07	0.05	0.04
	<i>1</i>	0.31	0.40	0.53	0.61	0.64
<i>0.20</i>	<i>0</i>	0.18	0.17	0.14	0.11	0.09
	<i>1</i>	0.50	0.60	0.71	0.78	0.80
<i>0.30</i>	<i>0</i>	0.27	0.26	0.22	0.17	0.15
	<i>1</i>	0.63	0.72	0.81	0.86	0.87
<i>0.40</i>	<i>0</i>	0.37	0.35	0.30	0.25	0.22
	<i>1</i>	0.73	0.80	0.87	0.90	0.91
<i>0.50</i>	<i>0</i>	0.47	0.45	0.39	0.33	0.29
	<i>1</i>	0.80	0.86	0.91	0.93	0.94
<i>0.60</i>	<i>0</i>	0.57	0.55	0.49	0.43	0.38
	<i>1</i>	0.86	0.90	0.94	0.95	0.96

<i>PSP</i>	<i>Replication Success</i>	<i>Power</i>				
		<i>0.20</i>	<i>0.30</i>	<i>0.50</i>	<i>0.70</i>	<i>0.80</i>
<i>Adversarial Bias = 0.25</i>						
<i>0.70</i>	<i>0</i>	0.90	0.93	0.96	0.97	0.97
	<i>1</i>	0.78	0.76	0.72	0.66	0.62
<i>0.80</i>	<i>0</i>	0.94	0.96	0.98	0.98	0.98
	<i>1</i>	0.89	0.88	0.85	0.82	0.79
<i>0.90</i>	<i>0</i>	0.97	0.98	0.99	0.99	0.99
	<i>1</i>	0.09	0.08	0.07	0.05	0.04

NOTE: *PSP* is the probability that the relationship being studied actually exists following the results from an original, published study. *Power* is the probability that the replication will produce a statistically significant estimate given that a relationship actually exists. *Adversarial Bias* represents the decreased share of significant estimates that appear in replication studies in the published literature due to a bias against confirming original results. The values in the table represent the updated *Post-Study Probability (Relationship Exists)* following a successful (*Replication Success* = 1) and unsuccessful (*Replication Success* = 0) replication, calculated using Equations (7) and (8) in the text, respectively. Values can be checked, and alternative parameter values investigated, using the spreadsheet posted at: <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FMANV6Y>.

FIGURE 1
Six Different Types of Replications

<i>Measurement and/or Analysis</i>	<i>Source of Data</i>		
	<i>Same dataset</i>	<i>Same population</i>	<i>Different population</i>
<i>Same</i>	(1) <i>Reproduction</i>	(3) <i>Repetition</i>	(4) <i>Extension</i>
<i>Different</i>	(2) <i>Robustness Analysis – Same Dataset</i>	(5) <i>Robustness Analysis – Same Population</i>	(6) <i>Robustness Analysis – Different Population</i>

Source: Reed (2017)