Novelty vs. Replicability: Virtues and Vices in the Reward System of Science

Felipe Romero
Tilburg University[*][†]

March 30, 2017

Abstract

The reward system of science is the priority rule (Merton, 1957). The first scientist making a new discovery is rewarded with prestige while second runners get little or nothing. Strevens (2003, 2011), following Kitcher (1990), defends this reward system arguing that it incentivizes an efficient division of cognitive labor. I argue that this assessment depends on strong implicit assumptions about the replicability of findings. I question these assumptions based on meta-scientific evidence and argue that the priority rule systematically discourages replication. My analysis leads us to qualify Kitcher and Strevens' contention that a priority-based reward system is normatively desirable for science.

1 Introduction

The reward system of science is the *priority rule* scheme (Merton, 1957). In general, the first scientist making a discovery is rewarded with prestige and recognition while second runners get very little or nothing.[1] Arguably, the priority rule is functionally beneficial for science in various respects. According to economists of science, the priority rule incentivizes the production of scientific findings (Stephan, 2012). And foundational work in the social epistemology of science justifies the rationality of the priority rule in epistemic terms. In particular, Strevens (2003, 2011), following Kitcher (1990), argues that the priority rule incentivizes an efficient division of cognitive labor, allocating resources to (roughly) maximize science's payoff to society. On these grounds, Kitcher-Strevens's story suggests that a priority-based reward system is normatively adequate.

[*] Tilburg Center for Logic, Ethics and Philosophy of Science; f.romero@uvt.nl.

[†] I am grateful to Carl Craver, Frederick Eberhardt, Barton Moffatt, Anya Plutynski, Mark Povich, Jan Sprenger, and Roy Sorensen for comments on previous drafts. I also thank Teresa Ai, Thomas Boyer-Kassem, Matteo Colombo, John Doris, Julia Staffel, and Michael Strevens for helpful discussion. Previous versions of this work were presented at the American Philosophical Association Pacific Division meeting, San Francisco, CA, April 2016, and the Philosophy of Science Association Biennial Meeting, Atlanta, GA. November 2016. This research was partially supported by ERC Grant No. 640638, "Making Scientific Inferences More Objective".

[1] In practice, priority is not actual but perceived. In some cases, the work of the first scientist is not immediately known by the community, and another scientist who does a better job spreading the good news is perceived to be first.

In this paper, I dispute this contention. I argue that Kitcher-Strevens's story depends on strong implicit assumptions about the replicability of findings. Philosophers and scientists alike regard replicability as the gold standard of scientific findings in experimental science (Fisher, 1926; Popper, 2002; Heisenberg, 1975). I show that the interplay between the priority rule, statistical inference procedures, and publication practices favors novelty and discourages replication, hindering scientific confirmation.[2] This analysis leads us to qualify the alleged virtues of the reward system of science: while the reward system might be efficient in spreading out resources across novel projects, it is inefficient in allocating resources to produce well-tested theories. My analysis also offers a plausible explanation for the worrisome lack of replication in practice (Smith, 1970; Campbell and Jackson, 1979; Evanschitzky et al., 2007; Makel et al., 2012; Begley and Ellis, 2012; Makel and Plucker, 2014). As Longino remarks (Longino, 2015), the gap between the theory and reality of scientific self-correction has yet to be addressed by philosophers. This paper also takes some steps to study this gap.

In Section 1, I synthesize Kitcher and Strevens' assessments about the reward system of science. Then, in Section 2, I formulate the replicability principle using an example of a recent controversy about purported evidence of extrasensory perception; and then, in Section 3, I show how the priority rule creates a tension between novelty and replicability, which threatens the principle.

2 Virtues in the Reward System of Science: Novelty and Progress

Sociologist Robert Merton (1957) was probably the first to document the central role of the *priority rule* in the reward system of science.[3] Drawing from his work and also discussions about the priority rule in philosophy (Strevens, 2003, 2011) and economics (Stephan, 2012), I characterize the priority rule as follows:

(PR1) Only the first scientist (or research team) making a discovery is rewarded. Second runners get very little or nothing.

(PR2) Scientists' reward is not primarily money but *prestige*, i.e. peer recognition.[4]

(PR3) Prestige comes in different forms: eponymy[5], prizes (e.g., Nobel Prize, Fields Medal), and promotion.

Now, in addition to these characteristics, twentieth-century science adds an important characteristic:

(PR4) Scientists' establish priority for a finding via peer-reviewed publication. Prestige increases depending on the journal's prestige and citations to the publication.

[2] I'm concerned primarily with fields that rely on controlled experiments. In some research in e.g. anthropology or paleontology scientists can't replicate for practical reasons. However, hypothetically, such findings should replicate.

[3] Evidence comes from heated disputes over priority of findings, which have been common throughout the history of science. Merton documents classic examples (e.g., Newton vs Leibniz). Some recent examples are Montagnier vs Gallo over the discovery of HIV (a dispute that achieved political dimensions), Lauterbur vs Damadian over MRI, and MIT vs UC Berkeley over the genome-editing technology CRISPR.

[4] Although flattering, recognition by the general public is not the main reward. Scientists look primarily for recognition from others working in the same field (Hull, 1988, p.309)

[5] Look at the names of neurodegenerative diseases: Huntingon's, Parkinson's, Alzheimer's.

At first glance, this reward system might seem a capricious manifestation of human ego, not akin to the disinterested and rational scientific spirit. Nonetheless, some authors have exalted its functional significance. I summarize the functional role of the priority rule in two alleged virtues.

2.1 Virtue 1: The Priority Rule Incentivizes the Production of Novel Scientific Findings

Scientific findings are traditionally regarded as *public goods*. Once published, researchers cannot sell them and everybody else can benefit from them. This is a problem insofar as a market economy does not provide incentives to produce them (Stephan, 2012, p.6). Arguably, a priority-based reward system has solved this problem (Stephan, 2012). The scientist that makes a discovery appropriates it when her peers, and specially her competitors, give her credit for being the first making the discovery. Prestige is the currency, which incentivizes scientists to produce findings that otherwise they would be less likely to produce and share. There could be monetary rewards, but they are derived from priority. For instance, in the 20th century the patent system motivates research with commercial potential (Scotchmer, 2004). Still, patents depend on priority.

2.2 Virtue 2: The Priority Rule Incentivizes the Division of Cognitive Labor

The problem of the division of cognitive labor is "how to best allocate resources among different possible scientific projects" (Strevens, 2011). Kitcher states the problem as follows (Kitcher, 1990, 1993): Imagine research programs that pursue the same goal (e.g., a vaccine). Some programs have higher potential (i.e., higher success probabilities) than others. For truth-driven scientists, the epistemically rational decision from their individual viewpoint is to join only higher-potential programs. But to maximize science's payoff, the community has to divide labor to explore lower-potential programs as well. After all, the history of science teaches that programs/theories that seem less likely to succeed sometimes turn out to be correct/true. The problem is that achieving this division of labor requires some scientists to join programs that from their individual perspectives would be epistemically irrational to join. More generally: there is a distribution of epistemic goals for individual scientists, different from their rational distribution of goals, that yields a higher probability that the community as a whole will reach its epistemic goals. Then, the problem is, how can cognitive labor be allocated in a way that is epistemically irrational for some scientists?

Kitcher uses a formal decision-theoretic model to show that the reward system of science solves this problem. However, the particular argument that concerns me here can be stated qualitatively: communities in which scientists are both credit-driven and truth-driven will come close to the desired division of cognitive labor. Assume that when a research program succeeds, credit is divided among its contributors. If a scientist is truth-driven only, then she will join primarily higher-potential programs. But if she is also credit-driven, then she will take into account how many other scientists are working on a program. If the higher-potential program is too populated, then she is expected to join a lower-potential program: in case of success, credit would be divided among fewer researchers.

Strevens' rejoinder is that the priority rule actually serves this end, and furthermore promotes an *efficient* allocation of labor among research programs, i.e. an allocation that (roughly) maximizes science's payoff to society (Strevens, 2003). The argument is as follows. In Strevens' analysis, scientific races have a *winner-contributes-all* character: "the first discovery brings great benefits to science or society, while the second discovery brings nothing at all" (Strevens, 2011). And recall (PR1), which Strevens calls the *winner-takes-all* character of the priority rule: scientists in the winner program get rewarded in proportion to their actual contribution to utility, while second-runners get nothing. Now, contrast two scenarios. A hypothetical scenario that rewards all programs in terms of their expected utility. And one in which (PR1) is the case. In both cases, scientists will divide themselves across different programs. However, more scientists will join higher-potential programs in the latter, because succeeding first matters. This makes the latter scenario more efficient. In Strevens' words, "scientists choose among research programs based on their prospects for earning scientific credit—status or reputation—and the reward system in science calibrates the allocation of credit so that the choices made lead to a distribution of labor that (roughly) maximizes the production of social good" (Strevens, 2011).

I will turn to discuss one core problem in this story.[6] Virtue 2 depends on the "winner-contributes-all" assumption. However, if this assumption is true, rewarded scientific findings ought to be for the most part replicable. To explain why this is a problem, I will first introduce some notions about replication using an example.

3 The Replicability Principle

A standard experimental paradigm for studying memory has the following form: (1) participants are presented a set of words, then (2) participants practice a random subset of those words, and then (3) participants take a free recall test. Unsurprisingly, participants are better at remembering practiced than non-practiced words (otherwise it would not make sense to study for exams). Now, in 2011, Cornell University's emeritus psychology professor Daryl Bem published an experiment inverting the order of (2) and (3). Participants had to take the test *before* practicing the words. His results show that participants were better at remembering words that they would practice *later* ($p = 0.002$, $d = 0.42$). Strange as it may sound, Bem's study seems to provide evidence for extrasensory perception (ESP). After two editors and four reviewers evaluated the paper, it was published in the leading social psychology journal, *The Journal of Personality and Social Psychology* (Bem, 2011).

Can we regard this finding as a scientific finding? Leaving aside a complete analysis of the term, by "scientific finding" I mean a (report of a) finding that is *truth-tracking* (but not necessarily true), and *trustworthy* to build upon. Does this finding have those qualities?

An immediate answer is to consider the replicability of the finding. Philosophers (Popper, 2002) and scientists (Fisher, 1926; Heisenberg, 1975) have strongly defended replicability as the gold standard of scientific findings. As Popper bluntly puts it, "non-reproducible single occurrences are of no significance to science" (Popper, 2002, p.64).[7] The relevant concept of replication here is often called "direct replication" in

[6] For discussion of other potential problems see Muldoon and Weisberg (2011).

[7] Popper defines a scientifically significant effect as "that which can be regularly reproduced by anyone who carries out the appropriate experiment in the way prescribed" (Popper, 2002, pp.23–24). Heisenberg, talking about Galileo's work, calls replicability "the essential basis for physics' success" and notes the role of replication in establishing agreement: "we can finally agree about [experiments'] results because we have learned that experiments carried out under precisely the same conditions do actually lead to the same results"

the literature: an experiment that mirrors an original experimental design in all factors that are purportedly causally responsible for the effect (Campbell and Jackson, 1979; Schmidt, 2009).[8] Using this concept, we can state a *replicability principle*. Suppose experiment E produces result F (e.g., some measurable quantity, such as an effect size $d = 0.42$ for ESP), then:

(RP) F is a scientific finding only if *in principle* a direct replication of E produces F.

I formulate RP in a minimal way. RP does not state that actual replication of E has taken place, only that in principle a direct replication should be successful.[9] Also, RP does not imply that F is scientifically important, or robust, or that it generalizes beyond its experimental conditions (although any F that purportedly has those properties should be directly replicable too).

But how can we ensure replicability? One possible answer is to assess E's methodological merits. This is what reviewers and editors do during peer review. In psychology, this means checking whether the authors did the appropriate controls, used adequate sample sizes, analyzed their data correctly, etc. If these checks are positive, then we can expect the experiment to replicate. I call this hypothetical replicability. Explicitly:

Hypothetical replicability: If E were directly replicated, then we would obtain F.

Establishing good methodology establishes hypothetical replicability. Now, consider Bem's ESP study again. Most researchers received Bem's findings with suspicion. The initial responses attempted to deny the paper's hypothetical replicability. It's difficult, however, to show what went wrong. The paper does not have methodological flaws according to the current standards for psychological research. Indeed, Bem is more careful than many researchers, and typical worries about psychology experiments don't apply to his study. For instance, the study rules out the common problem of small sample sizes, given that he tested more than 1000 participants. Also, the paper is arguably not a lucky shot, since he reported 9 experiments with similar findings. Now, from a classical statistical perspective, one can still try to speculate about possible methodological problems (Yarkoni, 2011): Bem's experiments have uneven sample sizes, which suggests data lumping and/or splitting; all p-values are close to the 0.05 threshold, which suggests selection bias; and, some features of the design suggest post-hoc hypothesizing. All these together arguably inflate the probability of false positives. Importantly, however, it's hard to confirm these suspicions. Not to mention that one could have the same suspicions about most published papers in experimental psychology, which raises the question of whether the suspicions undermine Bem's paper more than any other paper.

Another reaction is to try to argue that Bem's research evidences that methodological standards in psychology are inappropriate. In particular, one could blame the statistics. Advocates of Bayesian statistics claim that classical statistics overestimates Bem's

(Heisenberg, 1975, p.55). Also, for statistician R.A. Fisher, experimentally established facts are those that rarely fail to replicate (Fisher, 1926, p.504).

[8] Direct replications are often contrasted with conceptual replications, designs that modify some aspects of the original experiment with the aim of assessing the generality of the phenomenon. See Cartwright (1991) for a similar distinction. It's controversial whether conceptual replications confirm findings (Pashler and Harris, 2012).

[9] The criteria for determining whether original experiment and replication produce the same F depend what parameters we compare (e.g., p-values, effect sizes, or confidence intervals) and the acceptable similarity ranges.

results as evidence for ESP (Wagenmakers et al., 2011; Rouder and Morey, 2011). Bem's and colleagues response raises a standard worry about Bayesian analysis. In Bayesian statistics, hypotheses should be tested against non-committal priors, and under what Bem and colleagues take to be the right non-committal prior, Bayesian analysis favors ESP (Bem et al., 2011).[10] This discussion shows, if anything, that there is space for disagreement about what constitutes a non-committal prior and how to apply Bayesian analysis to controversial findings.

The focus on methodological worries does not take us far in assessing Bem's findings. Furthermore, even if his findings are hypothetically replicable according to standard methodology, there are other possible sources of error beyond methodological flaws. This is true more generally and not only about his findings. Specifically, meta-scientific evidence in various fields suggests that many successful published findings result from questionable practices: selective reporting (Rosenthal, 1979), exclusion of data points (John et al., 2012; Bones, 2012), *p*-hacking (Simmons et al., 2011), HARKing (Kerr, 1998), fraud (Fanelli, 2009), and experimenter's bias.

These sources of error are not the sort that we can rule out as rare exceptions. And since the methodological checks don't reveal them, we can't be satisfied with hypothetical replicability. Now, we can't merely say that a finding such as Bem's is the result of such practices. Instead, if we want to rule out such possibilities of error, we have to attempt to replicate the experiment in question. Consider then another notion of replicability:

Actual replicability: E has been directly replicated, and we have obtained F.

Actual replicability is not sufficient to ensure that F is a scientific finding. Originators and replicators may share mis-calibrated instruments, in which case F would be false and still replicable. Also, actual replicability is not necessary either.[11] However, actual replications are particularly effective to increase trust in science under the right conditions (Simonsohn, 2015). The intuition is simple. If you wonder whether F is truth-tracking and trustworthy, a particularly convincing answer is that E has been successfully replicated. And while there are diminishing returns in replicating the same finding, when multiple independent investigators succeed in finding F, the possibility of F being one lucky shot or the result of fraud, decreases substantially.

Now, how does actual replicability square with the alleged virtues of the reward system of science?

4 Vices: Novelty vs Replicability

These considerations about the efficiency of the reward system of science (Virtue 2) depend on the "winner-contributes-all" assumption (Strevens, 2011). But is this assumption true? Given a finding F, a winner contributes all only if the replicability

[10] In Bayes factor analysis, researchers have to specify priors for H_0 and H_1. Wagenmakers et al's (2011) prior for H_1 diffuses the possible effect sizes over a range of possible values. Bem and colleagues argue that this range is unrealistic, because it considers as possible effects that are uncharacteristically large for psychological research. Bem and colleagues propose that if ESP is real, then with probability 0.9 the effect size is less than $d = 0.5$, a prior that is theoretically informed by the knowledge of typical effects in psychology. Under such a prior, they show that Bayesian analysis of Bem's data favors ESP (Bem et al., 2011). To be clear, the disagreement is not about the prior probability of H_1, but about the particular form of H_1.

[11] Some episodes in the history of science show that findings have been established without direct replication (Chen, 1994). More work is necessary to specify the circumstances under which this works.

principle (RP) is true of F. However, the interplay between the reward system, publication practices, and statistical inference procedures creates particularly inappropriate conditions for original proposers of F and third parties to ensure RP. I will now discuss these two vices.

4.1 Vice 1: The Priority Rule Systematically Rewards Findings Regardless of their Replicability

Recall (PR4): in the twentieth century, the peer review process determines priority (and prestige) over a finding. However, as I discussed in the previous section, the peer review process is useful to uncover methodological flaws in an experiment, which establishes hypothetical replicability. But what about actual replicability? Nothing done in the peer review process can establish actual replicability. In particular, reviewers of experimental papers don't attempt to replicate the work they are refereeing (at least not as part of their refereeing duties). Hence, establishing priority (and obtaining prestige) for a finding does not imply that the finding is actually replicable. Using Strevens' terms, winners take all but may not contribute all.

Meta-scientific evidence confirms this theoretical argument. In psychology, a recent attempt to assess the replicability of published research suggests that only around 39% of published findings directly replicate (Open Science Collaboration, 2015). Replicability of published cancer research is also low. Biotechnology firms often try to capitalize on preclinical studies, turning promising findings into drugs. With this end in mind, in the late 2000's Amgen Inc. identified 53 papers of preclinical cancer research deemed as "landmark" in the literature. Scientists at Amgen attempted to directly replicate the experiments, and even contacted the original researchers to perform direct replications under their supervision. They report, however, that they could only replicate 11% of the experiments (Begley and Ellis, 2012). Now, Amgen's report also evidences that authors often get their rewards regardless of the replicability of (and even controversies about) their findings: the report traces citations of the papers in the study, and shows that citation counts to the non-replicable and replicable papers are roughly the same; and the non-replicable papers have produced secondary publications that do not intend to confirm or falsify them (Begley and Ellis, 2012, p.532).

Now, someone might object that ensuring the replicability principle is not a function of the reward system. This line of reasoning does not make my argument unsound but might make the conclusion seem less like a vice. However, this line of reasoning raises a question for the objector: how does (or should) science ensure the replicability principle?

There are two possible responses, not mutually exclusive but equally unsatisfactory. First, the objector might respond that the original researcher should ensure that her findings are replicable *before* attempting to publish them. Nonetheless, if the reward system rewards findings regardless of their replicability, then scientists cannot be rationally expected to ensure replicability for most of their findings. After all, ensuring replicability is more costly than not ensuring it. Second, and more important, successful replications by the same proposer of a finding often are not convincing for the community, and should not be. (Recall Bem's nine experiments.) Such replications could be subject to systematic biases (e.g., experimenter expectancies and miscalibrated instruments).

The second possible response is that *others* in the community (should) ensure the replicability principle. That is, the objector could argue, actual replication takes place anyway, because potential users of a finding attempt to replicate it after it has been published. Thus, in the long run, findings that survive will be replicable. This response, however, depends on users of findings having incentives to replicate. Consider the next vice.

4.2 Vice 2: The Priority Rule Systematically Disincentivizes Direct Replication of Others' Findings

(PR1) and (PR4) create a problem for replication of others' work. There are two possible cases when someone tries to replicate an experiment. If the replication succeeds, then it is not expected to receive high rewards because it is not a novel result, even if it contributes epistemically (e.g., the replication might provide a more accurate estimation of an effect size). If the replication fails, then, in theory, it can be expected to be rewarded: the replicator would be the first to prove the original experimenter wrong. Indeed, following Kitcher (1993, pp.339–340), one may think that when the original finding comes from a famous researcher, there are high incentives to replicate and high rewards on a failed replication.

In practice, however, the credit-driven scientist who wants publications has reasons to do novel research instead of replication work, because publishing replication research is very difficult. In classical statistics, a negative result (understood as a failure to reject a null hypothesis) is typically inconclusive (Simmons et al., 2011). Hence, a failed replication qua negative result is difficult to publish in high profile journals. To get a publication, the credit-driven scientist not only needs to fail to replicate the finding, but also make the argument that the null hypothesis is true. And the rules to make this sort of inference are controversial (Machery, 2012).

It has also been known for a while that the editorial system is biased against replication research. Most editors regard showing novel findings to be more important (Neuliep and Crandall, 1990, 1993). Negative results are disappearing in most disciplines (Fanelli, 2012, 2010). Published direct replication attempts in the social sciences are less than 1% of published findings (Makel et al., 2012; Makel and Plucker, 2014; Evanschitzky et al., 2007). And, at least in psychology, such attempts are mostly made by the same team that produced the original finding (Makel et al., 2012). Although not conclusively, this questions whether the incentives to replicate others' findings are practically effective.

Bem's ESP controversy illustrates these vices. After Bem's publication, Ritchie et al. (2012) performed a three-lab replication study, which turned out to be unsuccessful. They struggled to publish their article (Ritchie and French, 2012). The same journal that published the original article —*The Journal for Personality and Social Psychology* (JPSP)—rejected the replication study on the basis that the journal does not accept direct replications (either successful or unsuccessful). *Psychological Science* rejected the paper on the same grounds (Winerman, 2013). Another replication study suggesting no evidence for ESP had more luck, and it was published in JPSP (Galak et al., 2012). Unlike Ritchie et al.'s study, Galak et al.'s study is not merely a direct replication since it provides a meta-analysis of replication attempts of Bem's experiments from different sources, some of them unpublished, which arguably made the article more likely to be published under the journal's policies.

The lessons from this story go beyond ESP. The perhaps surprising fact is that this situation is the same for many findings that are less controversial. Furthermore, the situation is worse when scarce resources and career pressures exacerbate the vices. Such findings could turn out to be false, but without direct replication attempts of the right sort, it will be hard to know.

5 Conclusion

The thesis that the reward system of science incentivizes an *efficient* division of cognitive labor is too optimistic. This thesis assumes that the scientific community in most opportunities produces and rewards replicable findings. This assumption is not unreasonable in the contexts that Merton had in mind when he wrote about the priority rule (e.g., historical cases of discovery in physics and mathematics). Nonetheless, in the context of contemporary experimental science, the assumption is highly questionable. Theoretically, a priority-based system can at most ensure hypothetical replicability but not actual replicability. And in practice, many rewarded findings are not replicable at all.

We need further work developing recommendations to address the vices in the reward system of science. We need to design scientific institutions that divide cognitive labor in a way that produces both new discoveries and well-tested theories. Such recommendations should consider that credit-driven scientists may not be suited for the latter task. We should also consider that different scientific fields have different notions of success, and therefore might need different reward systems.

More generally, there is also a meta-philosophical upshot: Kitcher and Strevens' work has inspired highly abstract formal studies on the social epistemology of science. If the aim is to say something relevant about science (and not only to develop formal models), these studies should take statistical inference practices and meta-scientific evidence more seriously.

References

Begley, C. Glenn and Lee M. Ellis (2012). Drug Development: Raise Standards for Preclinical Cancer Research. *Nature 483*, 531–533.

Bem, Daryl J. (2011). Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect. *Journal of Personality and Social Psychology 100*(3), 407–425.

Bem, Daryl J., Jessica Utts, and Wesley O. Johnson (2011). Must Psychologists Change the Way They Analyze Their Data? *Journal of Personality and Social Psychology 101*(4), 716–719.

Bones, Arina K. (2012). We Knew the Future All Along: Scientific Hypothesizing is Much More Accurate Than Other Forms of Precognition-A Satire in One Part. *Perspectives on Psychological Science 7*(3), 307–309.

Campbell, Keith E. and Thomas T. Jackson (1979). The Role and Need for Replication Research in Social Psychology. *Replications in Social Psychology 1*(1), 3–14.

Cartwright, Nancy (1991). Replicability, Reproducibility, and Robustness: Comments on Harry Collins. *History of Political Economy 23*(1), 143–155.

Chen, Xiang (1994). The Rule of Reproducibility and Its Applications in Experiment Appraisal. *Synthese 99*(1), 87–109.

Evanschitzky, Heiner, Carsten Baumgarth, Raymond Hubbard, and J. Scott Armstrong (2007). Replication Research's Disturbing Trend. *Journal of Business Research 60*(4), 411–415.

Fanelli, Daniele (2009). How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. *PLoS ONE 4*(5), 1–11.

Fanelli, Daniele (2010). Positive Results Increase Down the Hierarchy of the Sciences. *PLoS ONE 5*(4), e10068.

Fanelli, Daniele (2012). Negative Results are Disappearing from Most Disciplines and Countries. *Scientometrics 90*(3), 891–904.

Fisher, Ronald A. (1926). The Arrangement of Field Experiments. *Journal of the Ministry of Agriculture 33*, 503–513.

Galak, Jeff, Robyn A. LeBoeuf, Leif D. Nelson, and Joseph P. Simmons (2012). Correcting the Past: Failures to Replicate Psi. *Journal of Personality and Social Psychology 103*(6), 933–948.

Heisenberg, Werner (1975). The Great Tradition: End of an Epoch? *Encounter 44*(3), 52–58.

Hull, D.L. (1988). *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*. Science and Its Conceptual Foundations S. University of Chicago Press.

John, Leslie K., George Loewenstein, and Drazen Prelec (2012). Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling. *Psychological Science 23*(5), 524–532.

Kerr, Norbert L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review 2*(3), 196–217.

Kitcher, Philip (1990). The Division of Cognitive Labor. *Journal of Philosophy 87*(1), 5–22.

Kitcher, Philip (1993). *The Advancement of Science: Science Without Legend, Objectivity Without Illusions*. Oxford University Press.

Longino, Helen (2015). The Social Dimensions of Scientific Knowledge. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2015 ed.).

Machery, Edouard (2012). Power and Negative Results. *Philosophy of Science 79*(5), 808–820.

Makel, Matthew C. and Jonathan A. Plucker (2014). Facts are More Important than Novelty: Replication in the Education Sciences. *Educational Researcher 43*(6), 304–316.

Makel, Matthew C., Jonathan A. Plucker, and Boyd Hegarty (2012). Replications in Psychology Research: How Often Do They Really Occur? *Perspectives on Psychological Science 7*(6), 537–542.

Merton, Robert K. (1957). Priorities in Scientific Discovery: A Chapter in the Sociology of Science. *American Sociological Review 22*(6), 635–659.

Muldoon, Ryan and Michael Weisberg (2011). Robustness and Idealization in Models of Cognitive Labor. *Synthese 183*(2), 161–174.

Neuliep, James W. and Rick Crandall (1990). Editorial Bias Against Replication Research. *Journal of Social Behavior & Personality 5*(4), 85–90.

Neuliep, James W. and Rick Crandall (1993). Reviewer Bias Against Replication Research. *Journal of Social Behavior & Personality 8*(6), 21–29.

Open Science Collaboration (2015). Estimating the Reproducibility of Psychological Science. *Science 349*(6251), aac4716.

Pashler, Harold and Christine R. Harris (2012). Is the Replicability Crisis Overblown? Three Arguments Examined. *Perspectives on Psychological Science 7*(6), 531–536.

Popper, Karl R. ([1959] 2002). *The Logic of Scientific Discovery*. Classics Series. Routledge.

Ritchie, Stuart J., Richard Wiseman and Christopher C. French (2012). Replication, Replication, Replication. *The Psychologist 25*, 346–357.

Ritchie, Stuart J., Richard Wiseman, and Christopher C. French (2012, March). Failing the Future: Three Unsuccessful Attempts to Replicate Bem's 'Retroactive Facilitation of Recall' Effect. *PLoS ONE 7*(3), e33423.

Rosenthal, Robert (1979). The File Drawer Problem and Tolerance for Null Results. *Psychological Bulletin 86*(3), 638–641.

Rouder, Jeffrey N. and Richard D. Morey (2011). A Bayes Factor Meta-analysis of Bem's ESP Claim. *Psychonomic Bulletin & Review 18*(4), 682–689.

Schmidt, Stefan (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology 13*(2), 90–100.

Scotchmer, Suzanne (2004). *Innovation and Incentives*. MIT Press.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science 22*(11), 1359–1366.

Simonsohn, Uri (2015). Small Telescopes: Detectability and the Evaluation of Replication Results. *Psychological Science 26*(5), 559–569.

Smith, Nathaniel C. (1970). Replication Studies: A Neglected Aspect of Psychological Research. *American Psychologist 25*(10), 970–975.

Stephan, Paula E. (2012). *How Economics Shapes Science*. Harvard University Press.

Strevens, Michael (2003). The Role of the Priority Rule in Science. *Journal of Philosophy 100*(2), 55–79.

Strevens, Michael (2011). Economic Approaches to Understanding Scientific Norms. *Episteme 8*(2), 184–200.

Wagenmakers, Eric-Jan, Ruud Wetzels, Denny Borsboom, and Han L. J. van der Maas (2011). Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi. *Journal of Personality and Social Psychology 100*(3), 426–432.

Winerman, Lea (2013). Interesting Results: Can they be Replicated? *Monitor on Psychology 44*(2), 38–41.

Yarkoni, Tal (2011). The Psychology of Parapsychology, or Why Good Researchers Publishing Good Articles in Good Journals can Still get it Totally Wrong [Web log post]. *Retrieved June 3rd 2015 from http://tinyurl.com/694ycam.*