Who Should Do Replication Labor?

Felipe Romero[1][2]
University of Groningen

Abstract. Scientists, for the most part, want to get it right. However, the social structures that govern their work undermine that aim, and this leads to nonreplicable findings in many fields. Because the social structure of science is a decentralized system, it is difficult to intervene. In this article, I discuss how we might do so, focusing on *self-corrective-labor schemes* (i.e., ways of distributing replication efforts within the scientific community). First, I argue that we need to implement a scheme that makes replication work *outcome independent*, *systematic*, and *sustainable*. Second, I use these three criteria to evaluate extant proposals, which place the responsibility for replication on original researchers, consumers of their research, students, or many labs. Third, on the basis of a philosophical analysis of the reward system of science and the benefits of the division of cognitive labor, I propose a scheme that satisfies the criteria better: the professional scheme. This scheme has two main components. First, the scientific community is organized into two groups: *discovery researchers*, who produce new findings, and *confirmation researchers*, whose primary function is to do confirmation work (i.e., replication, reproduction, meta-analysis). Second, a distinct reward system is established for confirmation researchers so that their career advancement is separated from whether they obtain positive experimental results.

Keywords

replication, reproduction, incentives, scientific self-correction, philosophy of science

---

[1] Corresponding Author: Felipe Romero, Faculty of Philosophy, University of Groningen, Oude Boteringestraat 52, 9712 GL Groningen, The Netherlands E-mail: c.f.romero@rug.nl

[2] Acknowledgments: I am grateful to Mark Brandt, Joanne Chung, Matteo Colombo, Frederick Eberhardt, Anya Plutynski, Jan Sprenger, Jelte Wicherts, and three anonymous reviewers for comments on previous drafts. I also thank Teresa Ai, Carl Craver, John Doris, and Carlo Garofalo for helpful discussion. Previous versions of this work were presented at the Philosophy-Neuroscience-Psychology Colloquium, Washington University in St. Louis, St. Louis, MO, April 22, 2016; the annual meeting of the American Association for the Advancement of Science, Boston, MA, February 20, 2017; the Philosophy Colloquium at the University of Groningen, Groningen, The Netherlands, January 17, 2018; and the Philosophy Colloquium at the University of Utah, Salt Lake City, UT, January, 19, 2018. I thank those audiences for useful feedback as well.

Self-correction in the social and behavioral sciences faces three major problems. The first problem is epistemic: We need independent replications. Without them, we cannot rule out systematic error and increase our confidence in published findings. The second problem is sociological: Replication failures often lead to stalemates. Because replication attempts are not standard practice, reports of failed replications are often perceived as hostile attacks. Original and replication researchers question each other's qualifications and intentions instead of engaging in fruitful discussions about ideas. The third problem is economic: There are few material incentives to replicate studies. Despite the recent acknowledgment of the epistemic importance of replication, and despite changes in editorial policies in some journals, replication is still underrewarded second-class work relative to novel research. Can the social and behavioral sciences overcome these problems?

Addressing these problems, I suggest, requires intervening on the social structure of science, a difficult task given that science is a decentralized system.

Prominent intervention proposals range from multisite replication projects to educational replication exercises. Inspired by the philosophical literature on the division of cognitive labor (Kitcher, 1990; Strevens, 2003), I call these proposals *self-corrective-labor schemes*: specifications of roles, responsibilities, and communication protocols for scientific workers to organize their replication efforts. Philosophers think that modern science requires dividing cognitive labor (e.g., organizing scientists into different fields) because the complexity of modern science exceeds the cognitive capacities of individuals (Weisberg & Muldoon, 2009). Here I argue that, contrary to what we see in practice, scientific self-correction is also a process that requires such a division. Specifically, I present evaluation criteria that self-corrective-labor schemes should satisfy to solve the epistemic, sociological, and economic problems and thereby improve scientific self-correction. I then use these criteria to evaluate four schemes that have been proposed and defend a new proposal, which I call *the professional scheme*. This scheme divides the scientific community into two groups that operate under different incentive structures. One group focuses on producing new discoveries, and the other group supports the self-corrective process by doing primarily confirmatory research (i.e., replication, reproduction, meta-analysis, and theory criticism). Finally, I argue that one way of implementing the professional scheme is for stakeholders involved in the research process (from funding agencies to journals) to support *confirmation-research tracks* for professors at universities.

Framing the replicability discussion in terms of evaluation criteria that a self-corrective-labor scheme should satisfy provides a better philosophical understanding of the requirements that modern science must meet to be self-corrective. Traditional philosophical theories regard science as self-corrective (Peirce, 1901/1958; Reichenbach, 1938), but they are not informative about current practice. In addition, given that implementing self-corrective-labor schemes involves important cultural and structural changes, and that there are multiple proposals on the market, we need criteria to compare them and assess their potential efficacy, and this discussion is a step in that direction.

I have two caveats about the scope of this article. First, in response to the replicability crisis, some authors have forcefully defended statistical reforms. For instance, some advocate for banning *p* values (Trafimow & Marks, 2015) or replacing them with Bayesian statistics. I touch on this debate only tangentially. Although it is important to refine procedures for statistical inference, these refinements will likely fall short in increasing replicability unless flawed research and publication practices are also changed. Statistical tools can help us detect publication bias (Borenstein, Hedges, Higgins, & Rothstein, 2009; Light & Pillemer, 1984; Simonsohn, Nelson, & Simmons, 2014; van Aert, Wicherts, & van Assen, 2016). They can also help us detect and prevent errors in reporting (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016). However, they are not meant to substitute for replication.

The second caveat concerns the generalizability of my proposal. Many challenges to increasing replicability (e.g., defective incentive structures) are common to all research domains. However, there are likely no one-size-fits-all solutions. My proposal primarily concerns laboratory-based experimental social-science research with convenience samples (e.g., most of the research in social and cognitive psychology). Other kinds of research, for example, in developmental and clinical psychology, face additional challenges (Tackett et al., 2017), so my proposal would require further refinements to apply to them. At some points, I discuss such refinements, but only briefly.

1. Criteria for Evaluating Self-Corrective-Labor Schemes

What does it mean to say that science is self-corrective? A current approach to understanding this idea is based on the *parameter-estimation view* (Cumming, 2012; Schmidt, 1996), according to which the aim of scientific experimentation is to estimate the magnitude of parameters such as effect sizes and confidence intervals. In this view, a single experiment could in principle provide a valid estimation of, for example, an effect size. The problem is that we do not know that the estimate is valid because any single experiment could be subject to errors from a variety of sources. In theory, direct replications correct such errors in the long run. That is, as a series of replications of an experiment lengthens, the meta-analytic aggregation of their effect sizes should approach the true effect size (narrowing the confidence intervals). This is an attractive story. However, it is true only under highly idealized assumptions. First, this story fails in an environment with publication bias and systematic confirmation biases, and when scientists conduct low-power experiments (Button et al., 2013; Ioannidis, 2008; Ioannidis, Munafò, Fusar-Poli, Nosek, & David, 2014; Nuijten, van Assen, Veldkamp, & Wicherts, 2015; Romero, 2016). Second, given that the concept of direct replication is an ideal, approaching it in practice requires careful consideration of how to define the target effects and how to compare original and replication results (Brandt et al., 2014). Third, self-correction requires updating hypotheses and theories, in addition to estimating parameters accurately. Hence, repetition of previous experimental designs is insufficient. Researchers need to explore alternative designs (e.g., considering potential moderators, mediators, and alternative data analyses) and evaluate broader theoretical implications.

Consider a pessimistic illustration of a scheme that is still in place these days: A lab designs and conducts a series of experiments. The "unsuccessful" ones (i.e., those with results that are not statistically significant), some of which are direct replications of others, go into a private file drawer. The successful ones are selected and reported in a manuscript, along with an engaging post hoc narrative that connects them. The manuscript is published. Some time later, a second lab gets interested in the publication and aims to extend its findings. The researchers consider directly replicating one of the original experiments, but they hesitate because direct replication is not rewarded. They skip replication, persuaded by the various successes in the original research. Then, they design and conduct a series of conceptual replications. They select the successful ones, and the cycle repeats.

Most researchers (we hope) agree that we should replace this scheme. Nobody is explicitly responsible for checking that findings are replicable. Under this scheme, science does not self-correct. In this article, I present and compare five schemes that arguably do a better job. To make the comparison, I introduce three evaluation criteria that offer a principled way of assessing their theoretical self-corrective merits.

1.1. Outcome-independence criterion: replication labor has to be outcome independent

Would you trust a food company that says that their products are safe for human consumption and should not be subject to independent control by the Food and Drug Administration? Should we not treat scientific findings with as much care? Many fields have historically neglected independent replication. However, in light of the replication crisis, the need for it has become more apparent (Asendorpf et al., 2013; Makel, Plucker, & Hegarty, 2012; Tackett et al., 2017). What does "independent replication" mean in the context of social and behavioral science? Here I focus on the notion of *outcome independence*. Replications are outcome independent when the person conducting the replication research does not have a conflict of interest regarding the outcome of that research. That is, the researcher's own goals, projects, and prestige do not depend on the truth or falsity of the finding, and therefore the researcher does not have a strong preference for a specific outcome. According to this notion, replications by original researchers are not outcome independent. Yet most replication work in psychology is conducted by original researchers. Makel et al. (2012) reported that if a study is replicated at all, 52.9% of the time the same research team that published the original report (or an overlapping team) conducts the replication. Moreover, such replication attempts are more likely to replicate the original findings than are replication attempts conducted by other teams (Makel et al., 2012). Notice, however, that a replication attempt by another team is not necessarily outcome independent. Even if the original and replication teams are distinct, the latter could have conflicting interests if they share theoretical commitments or academic genealogy with the former.

The argument for outcome-independent replication research is that it is the most effective way of correcting (nonrandom) error and, at the same time, signaling to other researchers that the

literature can be trusted. More specifically, outcome-independent replication research corrects at least three types of error, as I discuss next.

1.1.1. Error due to questionable research practices.

A published (successful) study could have involved questionable research practices (QRPs; Bones, 2012; John, Loewenstein, & Prelec, 2012), *p*-hacking (Simmons, Nelson, & Simonsohn, 2011), or HARKing (hypothesizing after results are known; Kerr, 1998). Researchers may engage in QRPs and *p*-hacking when they strongly prefer specific outcomes. This can be the case both for an original experimenter who is strongly invested in establishing a finding and for a replication researcher who is invested in the falsity of a previous finding. If replication work is outcome independent, the researchers who conduct it are less likely to engage in these practices. This goal can be supported further with preregistration under the right conditions (see the discussion of preregistration later in this article).

1.1.2. Error due to unconscious confirmation or disconfirmation biases.

Even the most honest and careful scientist can fall prey to unconscious biases (Ioannidis et al., 2014; Nuzzo, 2015). In the case of replication work, original researchers could be subject to unconscious confirmation biases when they attempt to replicate their own work, and other researchers could be subject to unconscious disconfirmation biases. Nonetheless, when other researchers replicate an experiment, they can identify potential sources of bias that original researchers did not consider. As an example, consider the case of John Bargh's controversial priming experiment (Bargh, Chen, & Burrows, 1996), in which the dependent variable was walking speed after being presented with words related to the stereotype of elderly people. In the original experiment, time was measured manually with a chronometer. Two failed replications used infrared sensors instead (Doyen, Klein, Pichon, & Cleeremans, 2012; Pashler, Harris, & Coburn, 2011). Although other factors could explain the discrepancy, the insightful decision to use infrared sensors neutralized potential confirmation biases.

1.1.3. Error due to fraud.

According to survey data, fraud (understood as conscious fabrication, falsification, or modification of data) is rare (Fanelli, 2009). In contrast, QRPs are much more prevalent, subtle, and therefore worrisome as a contributor to nonreplicable research findings. Nonetheless, if fraud is not detected by statistical or social means, the errors that it introduces would still be corrected by outcome-independent replication work.

1.2. Systematicity criterion: replication labor has to be systematic

Replication failures often lead to stalemates. The original researcher shouts "positive!" the replication researcher responds with "false positive!" and then the original researcher responds with "false false positive!" The problem is not so much that the researchers disagree. After all, cutting-edge science is full of disagreements. However, when replication work is not standard

practice, disagreements after replication failure are often unfruitful, as they tend to focus on epistemically irrelevant factors (e.g., the other person's qualifications and alleged ill intent). In such cases, uncertainty about the truth of the finding persists in the community.

When all we have is one experiment and one replication failure, both the original and the replication researchers often have good reasons to cast doubt on the other's work. On the one hand, replication studies often use larger samples than original experiments. Moreover, if the replication study is properly powered and preregistered and the original experiment is not, then the former is likely more informative (Simonsohn, 2015). On the other hand, the original experimenter can sidestep the failure in several ways, for example, by attributing the failure to undiscovered moderators or lack of understanding about the mechanisms responsible for the phenomenon (Cesario, 2014). Another option is to suggest that the replication experiment was not executed correctly because the design is hard to export to other laboratories (Bissell, 2013). In response, the replication researcher can argue that science does not progress if finding an effect requires an incommunicable special touch. In more unfortunate, but common, scenarios, disagreement persists because the original experimenter questions the replication researcher's motives and interprets the replication attempt as a personal attack.[1] Under what conditions can these unfruitful disagreements be reduced?

Consider a scenario in which replication work is systematic: The same finding has been subject to multiple replication attempts, as opposed to only one. Systematic replication helps to overcome stalemates in two ways. First, it increases the reliability of estimates. All other things being equal, one original experiment and one replication experiment make equal contributions to the estimate of a parameter.[2] However, taken cumulatively, multiple replication experiments narrow confidence intervals. And more important, concerns about the three types of error just mentioned in connection with the outcome-independence criterion lose weight when it comes to systematic replication attempts because results of multiple replication experiments converge. Second, systematic replication research can reveal the effects of moderators. Moderating variables can lead original and replication experimenters to stalemates when both sides provide arguably reliable evidence. Systematic replication experiments can help us discover moderation effects by manipulating the values of potential moderators.

Note, however, that replication research needs to be both systematic and outcome independent to reduce unfruitful disagreement. Systematic replication research alone is not sufficient for obtaining valid estimates. Also, independent but nonsystematic replication experiments may not produce reliable estimates. Consider two scenarios. In the first scenario, 10 independent scientists attempt to replicate an experiment, and they pool the results of their 10 experiments at the end. In the second scenario, one scientist attempts 10 direct replications of the experiment. For both scenarios, assume equal sample sizes, materials, experience, weather, and so forth. In the absence of nonstatistical error (e.g., due to QRPs or fraud), valid estimates would be produced in both scenarios. That is, the meta-analytic aggregation would approach the true value of the parameter in question. However, in the presence of a nonstatistical error, the second scenario would not lead

to a valid estimate (and a body of literature consisting of such experiments would misleadingly signal validity.) Because in practice we do not know whether reported results of individual experiments reflect such errors, the first scenario ought to be preferred.

Note also that adequate power, although mandatory, does not confer the same benefits as systematic and outcome-independent replication work. Granted, all other things being equal, an experiment with a larger sample size produces a more reliable estimate than an experiment with a smaller sample size (i.e., the former yields a narrower confidence interval). Also, when *p* values determine publication, published experiments with small samples tend to overestimate effect sizes (Button et al., 2013; Ioannidis, 2008)—and these overestimations could have been prevented with large samples. Nonetheless, a large sample alone does not ensure the validity of an estimate because it does not rule out QRPs, bias, and fraud.

Why has systematic and outcome-independent replication work been neglected in psychology? We can partially understand why by comparing psychology with fields in which such work is more standard, such as medical research. Compared with false positives in psychological research, false positives in medical research tend to lead to higher downstream costs (e.g., human lives), so there is pressure to adopt stringent (and costly) confirmation standards. Additionally, given the need to compare alternative treatments, medical research has focused on estimating effect sizes accurately (which requires systematic replication work) more than psychological research has. Another potential explanation is that psychologists did not need systematic and outcome-independent replication research during the rise of experimental psychology, perhaps because, for years, they could detect many effects without it.

1.3. Sustainability criterion: replication labor has to be sustainable

Incentives for replication work are few (Koole & Lakens, 2012; Nosek, Spies, & Motyl, 2012). Sociologists, philosophers, and economists characterize the reward system of science as following the *priority rule* (Merton, 1957; Stephan, 2012; Strevens, 2003): Scientists' reward for their work is primarily the community's acknowledgment of their being first to make novel discoveries. Currently, scientists establish their priority via peer-reviewed publications and receive credit when colleagues cite and use their work. Under this reward system, the goal of increasing the frequency of replication research faces an immediate theoretical problem: It is primarily novel science that is rewarded, and replication research is not novel. This threatens independent replication work. Additionally, under this reward system, credit is divided: As the number of scientists exploring one theoretical idea or methodological approach (broadly construed) increases, other scientists are discouraged from exploring the same idea or approach. This threatens systematic replication work. In short, replication labor is sustainable when there are incentives in place that make replication research a standard practice.

In theory, the priority rule may provide incentives for replication work under certain conditions. A successful direct replication can be rewarded if the author supplements it with novel

information (e.g., meta-analyses or conceptual replications), or if the original design is "improved" (e.g., by adding controls). Arguably, however, once these modifications are introduced, the experiments look less like direct replications. For failed replications, the standards rise. Merely failing to obtain a statistically significant result does not merit publication. To claim novelty, the replication researcher has to make a convincing case that the original finding is a false positive (e.g., by using an enormous sample size and failing to detect the smallest effect that would be of interest). In such a case, the researcher can claim novelty for being the first to prove the original experimenter wrong. Notice, however, that this incentive is effective only when the target finding is well known. Hence, if this is the only incentive, very few findings will be checked. Also, this incentive is epistemically pernicious because it opens the door to disconfirmation biases.

More precisely, changing the incentives to make replication labor sustainable requires solving two problems. The first is the *venue problem*—the problem that the publication system does not reward replication research. Proposals to change incentives have focused on this problem (see Nosek et al., 2012, for a discussion). The assumption is that if scientists have venues for publishing replication experiments (successful or not), they will attempt them more. Such venues have had moderate success. For example, public online archives of replication attempts, such as the PsychFileDrawer Web site, are excellent means for retaining valuable information, but given that they are not fully acknowledged as publication outlets, they provide few incentives, as evidenced by the relatively small number of reports that researchers have posted on them. Having journals dedicated to publishing replication research, such as the journal *Replications in Social Psychology* (Campbell & Jackson, 1979), has not worked either. The lesson is that the prestige of the venue matters. A more promising model to solve the venue problem is opening the doors for replication reports in prestigious journals (Cooper, 2016; Simons, Holcombe, & Spellman, 2014; Vazire, 2015). Moreover, journals and editors could even demand replication research. For instance, in response to low rates of successful replication of early published findings on genotype-phenotype associations, editors of high-profile journals adopted higher submission standards, including that manuscripts should report independent replication studies ("Freely Associating," 1999; Kraft, Zeggini, & Ioannidis, 2009).

Changing the incentives also requires solving a second problem: the *career incentives problem*. Imagine a world in which all prestigious journals change their policies and start publishing direct replication experiments but everything else remains the same. In such a world, what would be the career prospects of a scientist who does only replication work? Given the priority rule, those prospects would very likely be low. Hiring, grant, and promotion committees would still find novel research more valuable than replication research. The replication researcher's work would still be considered second-class. In such a world, as in this one, at critical stages in scientists' careers, it would be rational for them not to attempt replications and instead focus on doing novel work. Indeed, the rational strategy to get more publications would still be to conduct many exploratory studies with small samples as opposed to fewer carefully designed and adequately

powered studies (Bakker, van Dijk, & Wicherts, 2012). The lesson is that changes in publication practices should be accompanied by changes in career incentives more broadly.

The lack of replication attempts has concerned behavioral scientists for decades (Loevinger, 1968; Smith, 1970). Still, replication has not become standard practice. History teaches us that the incentives have not been sustainable. Novelty still trumps replication, even if scientists understand the epistemic value of replication. A self-corrective-labor scheme has to align the incentives for conducting replication research with career incentives. Otherwise, the present worries about replicability will not generate sustainable changes.

2. Evaluating the Self-Corrective-Labor Schemes

A self-corrective scheme that satisfies the criteria in the previous section would protect the scientific community from individual biases, unfruitful disagreements, and incentive conflicts, and would create suitable conditions for the self-corrective process to work. But is there such a scheme? In this section, I first consider four prominent proposals (see Table 1 for a summary). The first scheme is the *producer scheme*. In this scheme, original researchers are required to replicate their work before publication (Cesario, 2014; Roediger, 2012). The second scheme is the *consumer scheme*, which requires researchers to conduct direct replications of findings they want to replicate conceptually (Pashler & Harris, 2012). The third scheme is the *student scheme*, which places the responsibility of replication labor on students in methods classes (Everett & Earp, 2015; Frank & Saxe, 2012; Standing, Grenier, Lane, Roberts, & Sykes, 2014). The fourth scheme is the *multisite scheme*, in which a group of labs identifies a target study, each lab attempts to replicate it, and a coordinating lab aggregates the data (Ebersole et al., 2016; Klein et al., 2014; Open Science Collaboration, 2012). I evaluate the merits of these schemes in light of the three criteria described in the previous section. However, I argue that they do not do justice to scientific self-correction and propose that the *professional scheme* (see Table 1) does a better job. At the end of this section, I discuss how preregistration might improve these five schemes and what the limits of such improvement are.

2.1. Producer scheme

Some authors (Cesario, 2014; Roediger, 2012) contend that one important source of low replicability of published findings, at least in social psychology, is that researchers are not critical enough about their own work. In Roediger's (2012) words, if researchers "twisted, bent, and hammered" (p. 27) their own effects, they would find their own mistakes and would not pollute the literature with nonreplicable data. His lesson is to "replicate your own work prior to publication [and] don't let others find out that you are wrong or that your work is tightly constrained by boundary conditions" (Roediger, 2012, p. 28). I call this way of working the producer scheme. In this scheme, replication labor is the original researcher's responsibility. Once the researcher has collected data that shows a nonnegligible effect size, he or she executes the experiment again, more thoroughly, and can submit the work for publication only if the

experiment is successful again. In a sense, multistudy articles, which have been dominant in many subfields in psychology, implement this scheme.[3]

Table 1.
Characteristics of the Five Self-Corrective-Labor Schemes

Scheme	Parties in charge of the replication labor	Criterion[a]		
		Outcome independence	Systematicity	Sustainability
Producer	Researchers who originally obtained the finding	×	✓	✓
Consumer	Consumers of the finding	×	×	✓
Student	Students in methods classes	✓	×	✓
Multisite	Groups of collaborating labs	✓	✓	×
Professional	A dedicated group of confirmation researchers	✓	✓	✓

[a]The entries in these columns indicate which of the three criteria each scheme satisfies.

This scheme can make replication labor systematic, but its biggest problem is that it does not make replication labor outcome independent. As an example, consider Daryl Bem's (2011) controversial precognition article. Bem presented nine experiments in support of precognition. If we consider only direct replications, the experiment that other researchers have discussed the most (i.e., the one on retroactive facilitation of recall) was conducted successfully twice. And if we count conceptual replications, then Bem's article presented nine replications of his finding. Here we have a scientist who worked under the producer scheme. Bem replicated his work prior to publication, multiple times and in different ways. Nonetheless, researchers still question his work (Rouder & Morey, 2011; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011), and the community at this point is far from accepting the finding.

In general, replicating one's work before publication could ensure more consistent estimates, but multiple replications by the same researcher can be subject to the same nonstatistical sources of error as the original study. Hence, it is expected that disagreements will persist.

The producer scheme can satisfy the sustainability criterion if the editorial system enforces it, so that researchers who conduct their own replication experiments are not at a competitive disadvantage. Conducting additional replications increases the time to publication, but researchers can still claim priority after replicating results and publishing them.

2.2. Consumer scheme

The consumer scheme places the responsibility of replication labor on consumers of findings (i.e., researchers who want to conduct research informed by an original experiment). This scheme is a response to the proliferation of conceptual replications and the striking lack of direct replications. According to Pashler and Harris (2012), "performing conceptual rather than direct replication attempts interacts insidiously with publication bias, opening the door to literatures that appear to confirm the reality of phenomena that in fact do not exist" (p. 531). That is, researchers trust other scientists too much and verify too little. The consumer scheme addresses this problem. A researcher who is interested in conceptually replicating a finding that another researcher has published should attempt to directly replicate it first. The slogan for this scheme is "no conceptual replication without direct replication." Journals could implement this slogan as a nonnegotiable condition for publication. Thus, published articles with conceptual replications would always report a successful direct replication.

This scheme constitutes an improvement in outcome independence over the producer scheme. First, it blocks the pernicious interaction between conceptual replications and publication bias. Second, researchers other than the original researchers conduct replications. Nonetheless, the consumer scheme does not fully satisfy outcome independence. A scientist who plans to conceptually replicate a finding has a conflict of interest regarding the replication's outcome: He or she wants the required direct replication to succeed.

Without changes to the publication system, this scheme partially satisfies the systematicity criterion. If the direct and conceptual replications both succeed, all the information becomes available, and this contributes to systematic replication. But if a replication fails, the consumer who attempted the replication does not try to publish a report on the failure, and this information loss affects other researchers. In particular, other consumers may attempt potentially useless direct and conceptual replications of the finding when they don't know about its previous direct replication failures. Hence, at best, this scheme prevents propagation of errors but does not correct them, and it could even lead to bias. In a system that publishes everything, this scheme is more effective (van Assen, van Aert, Nuijten, & Wicherts, 2014).

As is the case with the producer scheme, the consumer scheme can satisfy the sustainability criterion if the editorial system enforces it. Investigators who want to publish novel research have to pass the replication requirement, which incentivizes them to do replication work.

2.3. Student scheme

The student scheme makes students the main contributors to replication efforts (Everett & Earp, 2015; Frank & Saxe, 2012; Standing et al., 2014). The general procedure is that as part of their training (e.g., in a research-methods class), students are required to attempt a replication of an assigned study. They conduct the experiment under the supervision of an experienced teacher and report the outcome publicly. Under this scheme, researchers learn about the importance of

replication early in their careers, and, in the meantime, many replication experiments are conducted.

There are multiple possible implementation variants of this scheme (Standing et al., 2014). For instance, depending on the complexity of the target study, the students can be undergraduates or graduates. Different groups of students across different semesters can attempt to replicate the same target study, thus providing useful cumulative information (e.g., see Grahe et al., 2018). This could be a step toward satisfying the systematicity criterion. (Fully satisfying this criterion would require that professors or other students work on the aggregation of these results.) Students can report their findings in an online database, such as the PsychFileDrawer Web site. Publishing a replication study could be made a standard Ph.D. requirement in all accredited programs (Everett & Earp, 2015).[4]

The student scheme can satisfy the outcome-independence criterion. For instance, to eliminate potential conflicts of interest, students should choose (or be assigned) to replicate a study that their own Ph.D. research does not build on, so that they are not hoping their replication attempts will succeed. This would make the student scheme superior to the consumer scheme.

From the perspective of professional researchers, the student scheme is an appealing solution. But what about the students' perspective? Trying to replicate other researchers' work while working toward a Ph.D. requires an investment in time that puts the Ph.D. process at risk. As Spellman (2015) pointed out, sometimes students try and fail to do replications so many times that they do not have anything to show after a year of work. On top of that, they must worry about doing novel research. A solution is to treat replication work like mandatory military service, as Everett and Earp (2015) suggested. In their proposal, replication projects are a requirement in all accredited Ph.D. programs. If every student incurs the same costs, no student is at a competitive disadvantage. Thus, the student scheme can satisfy the sustainability criterion.

One concern about the student scheme is that students (especially undergraduates) could lack the training necessary to perform certain experiments (Dijksterhuis, 2013). I think this should not be as great a worry in experimental psychology as it could be in basic research in the life sciences (e.g., wet laboratory research involving animal models.) And Standing et al. (2014) reported not having this problem in their experience. Nonetheless, the fact that this concern has been raised has a sociological moral. Typically, disagreements after replication failure involve professional scientists. And even when the original researchers and replication researchers have similar seniority, original researchers tend to question the abilities of replication researchers and not acknowledge the results. An original researcher may have the same or even a worse reaction when facing a failure to replicate by a student. If the replication fails, the original researcher could complain about the student's alleged lack of expertise. To address this issue, the student scheme could incorporate multisite replication. This would require the right sort of coordination by teachers, but could reduce worries about a single student's lack of expertise. Nonetheless, this approach would not solve unfruitful disagreement completely, and would still be very taxing for

students. The root of the worries concerning the student scheme, is that in this scheme, replication is still regarded as second-class work.

2.4. Multisite scheme

In the multisite scheme, rather than conducting isolated replication attempts, researchers at various sites coordinate their efforts to attempt independent replications of the same finding. First, a coordinating site identifies the target study. Next, each participating site conducts a replication, from data collection to analysis, and the labs then pool their results and report a joint analysis. Some Many Labs replication projects have implemented this scheme (Ebersole et al., 2016; Klein et al., 2014; Open Science Collaboration, 2012) and have provided insights about the robustness of some findings. For instance, the 36 labs that participated in the 2012 Many Labs project (Klein et al., 2014) attempted to replicate four classic anchoring experiments (Jacowitz & Kahneman, 1995). Their aggregated results showed very large effect sizes (ds $\approx$ 2) for the four experiments and confirmed that anchoring is one of the most reliable effects ever discovered in psychology. The same project, however, failed to replicate a flag-priming effect and a currency-priming effect.

The main virtue of the multisite scheme is that it necessarily makes replication work systematic, which is useful to overcome unfruitful disputes. For example, contrast the reactions of the authors whose priming findings failed to be replicated in Klein et al.'s (2014) project with the stalemates discussed earlier. Yong (2013) reported that "Travis Carter . . . , who led the original flag-priming study, says that he is disappointed but trusts [the Many Labs] team wholeheartedly" (para. 10), and that "Eugene Caruso . . . , who led the original currency-priming study says, 'We should use this lack of replication to update our beliefs about the reliability and generalizability of this effect'" (para. 10). Although personality differences could explain these differences in reactions, it is also possible that heated, negative reactions are more likely when original researchers face a single failure to replicate their findings rather than a multisite failure. Additionally, given that multiple labs often operate in different contexts (e.g., they can obtain samples from different populations), the multisite scheme offers not only more accurate estimation of an effect but also information about its heterogeneity and therefore its generalizability.

A limitation of this scheme is that it may not satisfy outcome independence well. If participating labs are self-selected, they may have a special interest in seeing the findings replicated (or not). This limitation can be addressed by preregistering the study, requiring participating labs to join the replication project before the coordinating site discloses the target experiment, and assigning labs to experiments randomly if there are multiple target experiments. If these steps are taken, this scheme would do a better job at producing an unbiased consensus.

A more serious challenge for the multisite scheme concerns the sustainability criterion. The 2012 Many Labs report (Klein et al., 2014) has 51 authors. Under the current reward system, in which prestige is at best divided among participants in a project (and is preponderantly attributed to the

first author), engaging in this kind of projects is not profitable in the long term. These days, replication is a hot topic. But for how long will this be the case? We can make a theoretical prediction. Ironic as it seems, multisite replication projects are not outside the priority rule. Currently, it is appealing for scientists to invest some of their time in replication projects because such projects are novel. But the priority rule implies that once the hype about such projects passes, scientists will have significantly fewer incentives to engage in them. For this situation to change, the community needs to implement further structural changes.[5]

2.5. Professional scheme: satisfying the criteria

The professional scheme is my proposed self-corrective-labor scheme. It has two main features. First, there is a division of cognitive labor according to the stages of the research process. Today we accept that scientific communities thrive if scientists specialize and organize their work in structured institutions. Francis Bacon, a central theorist in the development of the experimental method in the 17th century, was perhaps the earliest advocate of this idea. Present-day scientific institutions divide cognitive labor primarily by discipline: Most scientists specialize in (and contribute to) clearly separated fields. But cognitive labor could be divided further, institutionalizing scientists' specialization in steps in the research process. Indeed, Bacon's *New Atlantis* (1627/2000) novel illustrates this kind of division. The novel describes a utopian scientific society that aims at expanding human knowledge and to do so breaks the research process down into tasks and assigns them to different groups. For instance, some workers (called "Merchants of Light") collect facts, others (called "Pioneers" or "Miners") design experiments, others (called "Inoculators") execute the experiments and report findings, and still others (called "Interpreters of Nature") generalize the findings.[6]

The professional scheme adopts this kind of division. Researchers in the community are divided into two groups that specialize in different steps of the research process. For one group, *discovery researchers*, the main responsibility is to produce new discoveries. For the other group, *confirmation researchers*, the main responsibility is to support the self-corrective process of the community. This support involves primarily doing replication work (conducting direct replications and close replications that vary moderators and explore mediators, incorporate stricter controls, and test effects cross-culturally), but also conducting reproduction work (e.g., reanalyzing data, verifying hypotheses over existing data sets, conducting alternative statistical analyses), performing meta-analyses, and critiquing theory.

Methodologists and philosophers of science distinguish between exploratory and confirmatory research.[7] Hence, it is worth clarifying how these terms relate to the professional scheme. Exploratory research can be defined as research that looks for patterns (which might help to generate new hypotheses), whereas confirmatory research is research that tests predefined hypotheses or models. Although there are clear cases of each kind of research, there is also research that has both exploratory and confirmatory aspects (Wright, 2017). The point of the professional scheme is not to constrain discovery researchers to do only exploratory research.

They may well engage in confirmatory research, and to projects in between. Rather, the professional scheme is a system in which some scientists' responsibility is to adopt a purely confirmatory mode about *other* scientists' published work for error-control purposes. For instance, a project that has both exploratory and confirmatory aims for a discovery researcher can be later approached by a confirmation researcher from a confirmatory mode.

The second feature of the professional scheme is that it institutionalizes the division of labor because the two groups operate under different incentive structures. On the one hand, discovery researchers operate under the novelty-based economy that governs science in academia. Their goal is to produce new knowledge, their professional career depends mostly on the prestige they obtain, and any financial reward is derived from that prestige. On the other hand, confirmation researchers work under a service-based economy: They are compensated for the quality of their confirmation efforts, and their goal is not to get credit by publishing new discoveries.

The professional scheme builds on the insights of the other four schemes and addresses their failures. Unlike the previous schemes, the professional scheme fully satisfies the outcome-independence and sustainability criteria. Sustainability is achieved by having distinct reward systems that support the division of labor. Confirmation researchers' service-based reward system incentivizes them to engage consistently in replication efforts. Note that both the division of labor and the separation of reward systems are necessary. If the reward systems for the two groups of researchers were the same, then the demands for novelty would trump replication work when competition is high, as occurs in current practice. Establishing separate reward systems sets clear expectations for both kinds of workers and breaks the pernicious association between novelty and career opportunities for confirmation researchers.

The professional scheme fully satisfies outcome independence because confirmation researchers' incentives are completely disconnected from the outcomes of their studies. Their rewards and career advancement depend on how good and efficient they are at confirmation tasks and not on devising novel phenomena. Hence, they do not need to be invested in the success of any theoretical agenda.

To satisfy systematicity, the professional scheme can adopt the procedures of the multisite scheme. (The professional scheme does not entail that replication research is systematic, but systematic replication is nonetheless necessary for the self-corrective process to work.) That is, confirmation researchers can coordinate replication projects, identifying target findings and recruiting other confirmation researchers to attempt the replications, pool their results, and report them.

Before discussing how to implement the professional scheme, I discuss how incorporation of preregistration can refine each of the five schemes.

2.6. Refinements: study registration

Psychologists have been increasingly interested in preregistration (i.e., uploading a time-stamped, noneditable plan for data collection and analysis to a public platform before conducting a study) and other open-science practices as interventions to increase the replicability of results. The self-corrective-labor schemes I have discussed can incorporate preregistration. In this section, I discuss how doing so improves the original schemes, as well as the limitations that remain.

One important general concern about preregistration and open-science practices centers on their range of applicability. These interventions require refinements if they are to be as effective outside the domain of hypothesis-testing studies involving easy-to-collect data as they are within that domain. In research with hard-to-collect data (e.g., in clinical or forensic psychology), necessary changes to sampling procedures are often unforeseeable (Tackett et al., 2017). Moreover, in longitudinal research (e.g., in developmental psychology), preregistered hypotheses and analysis plans may become obsolete because new findings or analysis tools may appear during data collection. In such cases, sticking to preregistered plans is not useful. We need more work to understand how different midway updates to experimental designs affect the epistemic import of studies (e.g., we need to better categorize kinds of research along the exploratory-confirmatory continuum), and we also need more practical guidelines for registration procedures that are tailored to the research domain.

But let us assume for the sake of the argument that authors preregister their plans correctly and rigorously. In this case, the primary virtue of preregistration is that it reduces error because it reduces the degrees of freedom that researchers have at different stages in a research project, from devising hypotheses to reporting results (see Wicherts et al., 2016, for a comprehensive categorization.) For this reason, if discovery researchers use preregistration, then the replicability of their findings would increase from the very beginning. Also, preregistration would benefit all schemes, because correcting errors arising from degrees of freedom is one of the goals of the outcome-independence criterion.

Nonetheless, in the producer and consumer schemes, preregistration does not fully achieve outcome independence because it does not entirely correct errors arising from unconscious confirmatory and disconfirmatory biases, and it does not prevent or correct errors due to fraud. First, although preregistering the experimental protocol for a replication study means that the protocol is fixed, aspects of the execution of the experiment could still be flawed or remain (unconsciously) underspecified, which would allow biases to creep in. For example, think of Bargh et al.'s (1996) elderly-priming study again. Assume that the procedure for measuring time (i.e., an experimenter with a chronometer) was decisive for the experiment's initial success. If so, a hypothetical preregistered replication by the same authors before publication would have likely stated the methods in the same way, and the results could have been the same. Additionally, even with preregistration, replications in the producer and consumer schemes can be subject to "hypothesis myopia" (Nuzzo, 2015, p. 183); that is, the researcher may focus on collecting

evidence that supports the hypothesis without considering alternatives. In other words, preregistration in these schemes guarantees that replication researchers follow the recipe, but not that the recipe is good.

Second, even with preregistration, researchers can commit fraud. Granted, fraud is the least worrisome contributor to the replicability crisis. However, if we want to prevent and correct errors introduced by fraud and thereby increase trustworthiness of the published record, the producer and consumer schemes, even with preregistration, are less suitable than schemes that incorporate independent researchers. Indeed, mandatory preregistration could incentivize preregistration fraud, such as preregistering a study after looking at the data while collecting it or fabricating data that fit preregistered hypotheses.

Preregistration would do a better job in improving outcome independence when experimenters are not invested in specific outcomes, that is, in the student scheme (when students' own research does not depend on their replication project), the multisite scheme (when the assignment of labs to experiments is blind and random to avoid self-selection), and the professional scheme.

Regarding the systematicity and sustainability criteria, all other things being equal, preregistration does not improve any scheme. Indeed, preregistration introduces a new worry regarding sustainability. Preregistration requires additional work, so encouraging it without changing the reward system puts the researcher who preregisters studies at a competitive disadvantage: Others may avoid the extra work with no penalty. Currently, some journals create an incentive by using badges (i.e., icons that recognize good practices) for articles that report preregistered studies. This is a small bonus. But one could hope that, with enough cultural change, badges would create an indirect incentive: If all your colleagues publish articles that have preregistration badges, then you will have to preregister your experiments too. Otherwise, your articles will look bad.

Additionally, preregistration does not imply any course of action after study completion. Hence, if the researcher is highly invested in a positive result and the results are negative, then there is still a chance of information loss, which hinders self-correction in the long run. In an ideal world, the researcher in this situation would write up and share (and perhaps try to publish) a postreplication report. However, both in theory and in practice, preregistration does not incentivize reporting, nor does it penalize not reporting. For instance, results of at least 33% of registered clinical trials are not reported after the studies' completion (Chen et al., 2016).

A refinement that addresses some of the limitations of preregistration is to connect it to the publication system. The *Registered Reports* publication model (Chambers, 2013) provides such a connection. In this model, the researcher submits a research proposal (including experimental design and analysis plan) to a journal before data collection, the editors and reviewers evaluate it, and if it is methodologically sound (or revised so that it becomes sound), they give it an in-principle acceptance. This means that if the researcher executes the proposed experiment

correctly, then the report will be published regardless of the experimental outcome. This model rewards good methodology and makes expected outcomes less relevant and therefore less pernicious as an incentive. Also, given that accepted proposals incorporate critical editorial feedback, this model can further constrain researcher degrees of freedom in comparison with plain preregistration and reduce hypothesis myopia. The *Registered Replication Reports* model (Simons et al., 2014) combines this publication model with multisite work.

Schemes that incorporate Registered Reports would be better for controlling bias compared with schemes that incorporate only preregistration. However, they would not satisfy outcome independence when experimenters are self-selected because the experimenters would still conduct the experiments for which they have confirmation or disconfirmation interests and therefore could still make (unconscious) decisions that would bias the literature. In particular, an in-principle acceptance does not imply that the researcher will submit the final results to the journal. If the results are not consistent with the researcher's broader career interests, the researcher may withdraw the study and not submit the report for publication.

Finally, the success of preregistration with regard to sustainability is partial. If many prestigious journals require (and not merely allow) Registered Reports, then confirmation researchers would not have to worry about the venue problem. However, the career incentives problem would remain (see the earlier discussion of the student scheme).

In short, preregistration, when done correctly, reduces QRPs but does not imply good designs, nor does it guarantee reporting, and the decision not to report can introduce biases in the literature. Registered Reports do a better job addressing these problems: Designs reflect editors' and reviewers' feedback, and in-principle acceptances remove the dependence between outcome and publication. This model contributes to approaching outcome independence, although there is still the possibility of error due to self-selection (which may introduce biases) and fraud. These and additional limitations concerning the systematicity and sustainability criteria can be addressed if Registered Reports are incorporated in the right schemes. Schemes that remove self-selection (i.e., the student and multisite schemes with random assignment to studies and the professional scheme) would come closer to meeting the criterion of outcome independence. Schemes that incorporate multisite work (i.e., the multisite scheme and the professional scheme) would guarantee generalizability across labs. And schemes that address the career incentives problem (i.e., the professional scheme) would ensure sustainability.

3. How to Implement the Professional Scheme

The previous section established a theoretical ideal. The professional scheme separates reward systems for novel and confirmation research to solve the career incentives problem and make replication labor sustainable. At this point, the problems of the philosopher end, but those of the science policymaker begin. It is possible to imagine different approaches for implementing this scheme in practice. In this section, I discuss one of these approaches: establishing a system that

supports *confirmation-research-track* positions for professors at universities. To create such a system, it is necessary to align interventions at the levels of different stakeholders involved in the research process (see Table 2 for a summary).

Table 2.
Implementing the Professional Scheme: Interventions for Each Level of Stakeholders

Stakeholder	Intervention
Universities and departments	Create confirmation-research-track positions for professors Tailor tenure and promotion guidelines to acknowledge replication research
Governmental and private funders	Allocate steady funding for confirmation research Require multisite work for confirmation projects Standardize guidelines for evaluating confirmation-research proposals
Journals and editors	Publish multisite Registered Reports Publish reproducibility reports
Professional societies	Identify what is worth replicating in each subfield Define subfield-specific confirmation guidelines

3.1. Universities and departments

3.1.1. Create confirmation-research-track positions for professors.

Currently, the responsibilities of university professors in most fields lie along a research-teaching spectrum: There are pure research and pure teaching positions and many hybrid research-teaching positions in which time is allocated to each activity depending on the university's focus. To implement the professional scheme, I propose that departments segment the type of research expected from professors, creating (or allowing) positions in which confirmation research for error-control purposes is the main responsibility.

Scientists in these positions should have a variety of experimental and analytic skills. On the experimental side, confirmation researchers should be skillful at identifying confounds, assessing the generalizability of findings, and optimizing existing experimental protocols creatively. On the analytic side, confirmation researchers should be versed in alternative tools for statistical analysis to evaluate the reproducibility of findings, as well as meta-analytic tools to assess bodies of evidence more comprehensively. Additionally, given that replication work is resource intensive, confirmation researchers should be skillful at creating and sustaining effective collaborations to share resources with other confirmation labs. (This is particularly necessary for subfields that use

hard-to-collect data.) Such collaborations could involve students to make multisite work more resource efficient (keeping in mind the caveats of the student scheme discussed earlier). For example, when the difficulties of a study involve design more than execution (e.g., survey research and online experiments), a confirmation researcher could coordinate students to execute the study. Similarly, Ph.D. students in scientist-practitioner programs could contribute to the confirmation of findings that inform practitioners in their subfield.

This profile of necessary skills should be further adjusted according to the confirmation tasks that would be most useful for a given subfield. In subfields that rely on secondary data analysis (e.g., developmental psychology), confirmation researchers would need creativity to, for example, find alternative data sets and would need to be skillful at running different types of models on published data.

What would make confirmation researchers and discovery researchers different is not fundamentally their skills, however. The two groups would need similar training in Ph.D. programs. Indeed, ideally, there would be substantial overlap in the skill sets of the groups. Confirmation researchers would need to produce creative work, and discovery researchers would need to improve their experimental rigor.

Fundamentally, two factors would make the work of confirmation researchers and discovery researchers different. First, each type of scientist would have a distinct interest when approaching a finding. Discovery researchers would be after theoretical innovation, which may come at the expense of reliability. On the other hand, confirmation researchers would be primarily after findings' reliability, but would remain neutral about how innovative a finding is. The interests of the two groups would be distinct because they would derive from different reward systems. Second, and more important, the two groups would work with findings that are different in origin. Discovery researchers would be concerned about producing their own findings (although this would not exclude the possibility that they could engage in confirmation work). Confirmation researchers, on the other hand, would work primarily with other people's findings.

One could worry about the costs involved in establishing confirmation-research tracks, as this would require creating a new infrastructure and new administrative procedures at universities. And one could wonder why it is not sufficient to just fund confirmatory research more, so that researchers have incentives to do it? Although it is indeed indispensable to create funding structures to support confirmation research (as I explain later), such an intervention falls short in meeting the sustainability criterion. Making funds for confirmation research available and accessible would still leave confirmation efforts as an *option.* If no one is explicitly responsible for confirmation work, it could be neglected—and this is particularly likely if scientists are strongly required to produce novel research. On the other hand, establishing confirmation-research tracks would create clear expectations for conducting such work: For confirmation research scientists, confirmation would be not an option but a requirement.

A confirmation-research track could be of either of two types. In a *pure* confirmation-research track, all of a scientist's research time would be devoted to confirmation work. Less radically, in a *hybrid* track researchers would allocate a fraction of their time to confirmation efforts and the rest to exploratory research.

Pure confirmation-research positions would satisfy the outcome-independence criterion better than hybrid positions. Hybrid positions would not remove conflicting interests as well as pure positions because scientists in the former would operate simultaneously under two reward systems. These hybrid positions, however, would have a sociological advantage because they would not be perceived as making such a radical departure from the current system. Hence, hybrid positions could be useful during a transition period toward pure positions.

How feasible would it be to create confirmation-research tracks? Universities already allocate professors' time to different tasks (i.e., research, teaching, and administration). However, there have not yet been initiatives to create confirmation-research tracks for professors. One related precedent is that principal investigators have created positions in which Ph.D. candidates are primarily expected to carry out confirmation research.[8] Students trained in these positions would have the right profile to become professional confirmation researchers after graduation. Although this initiative may not become standard, it illustrates the possibility of a system in which some scientists are primarily dedicated to confirmation efforts.

3.1.2. Tailor tenure and promotion guidelines to acknowledge replication research.

Tenure and promotion requirements in psychology departments typically include criteria such as productivity, visibility, and originality. The criterion of originality as traditionally understood, of course, should be given less weight for confirmation researchers. However, the other guidelines should still apply. When assessing productivity and visibility, promotion committees should acknowledge that publishing replication research is hard. (Changing this situation will require additional interventions in the publication system, as I discuss later.) Confirmation researchers and their institutions should establish agreements to develop coherent research programs, and success in these endeavors should be an additional criterion for making tenure and promotion decisions.

Assessing confirmation researchers' performance will require assessing the quality of their work. I propose that high-quality confirmation work focuses on target findings whose status is highly informative for the community. Such findings have some (or all) of the following characteristics: They have not been well tested already (e.g., the effect has wide confidence intervals, the experiment had low statistical power, the experiment was not adequately controlled, and replications have not been attempted before), they have been influential in the literature, or they have societal relevance. (I return to this point when I discuss professional societies later in this section.) It will be important that confirmation outcomes (i.e., positive or negative) not be used as a metric to assess confirmation researchers' performance. Doing so would harm outcome

independence, as it would incentivize confirmation researchers to work on well-tested effects, which would allow them to produce consistent results (i.e., consistently confirmed or disconfirmed).

A related precedent is that an increasing number of postings of job openings have open-science requirements (i.e., in their job applications, candidates must explain how they address the replicability and reproducibility of their research). A next step would be to strengthen these requirements by asking candidates to provide a replicability and reproducibility agenda for a subarea of research more broadly conceived, which could lead to the hybrid-track model mentioned earlier. Also, some departments are increasingly incorporating quality metrics (as opposed to raw publication numbers) in their appointment and tenure-track criteria. These metrics include, for example, the number of studies with high a priori power, the number of preregistered studies, and the number of studies with open data and materials (see Schönbrodt, Heene, Maier, & Zehetleitner, 2015, for an example). At an institutional level, some universities have initiated efforts to implement university-wide research-quality requirements (e.g., see Dijstelbloem, Huisman, Miedema, & Mijnhardt, 2013).

3.2. Governmental and private funders

3.2.1. Allocate steady funding for confirmation research.

To make confirmation-research positions sustainable, the most important intervention would be to allocate funding directly and steadily to confirmation research. This intervention is the responsibility of funding agencies. Only in this way would it be possible to organize all the other stakeholders to maintain a system in which confirmation labor is a career option.

Funding agencies have started allocating resources to replication projects. In 2013, the Laura and John Arnold Foundation designated $1.3 million for replication studies on cancer biology findings ("Reproducibility Initiative," 2013). A few years later, the John Templeton Foundation funded a project to conduct independent preregistered replications of experimental studies on religion (John Templeton Foundation, 2017).

These fund-allocation precedents are steps toward the professional scheme. However, funding agencies should take further action. A project, by definition, has an end. In this sense, replication work should be seen not as a project, but as standard practice that does not end. To make replication work standard and ensure sustainability, part of the budget that funding agencies allocate to research every year should be reallocated consistently every year to fund replication efforts. A precedent along these lines has been set by the Netherlands Organisation for Scientific Research, which in 2016 launched a 3-year pilot program with €3 million to fund replication research (Netherlands Organisation for Scientific Research, 2016).

Future work should study epistemic trade-offs of different funding-allocation procedures. For instance, keeping resources constant, there is most likely a trade-off between increasing the

accuracy of estimation of already-reported effects and the discovery of new effects. A quantitative understanding of such a trade-off could inform allocation decisions.

3.2.2. Require multisite work for confirmation projects.

Funding agencies should not only support replication work but also require researchers applying for replication and confirmation grants to do multilab work. (Grant review committees could even recommend potential collaborators.) This intervention would enforce the systematicity criterion. Additionally, although this intervention is not meant primarily to increase statistical power, such an increase would be a welcome by-product. For projects that require hard-to-collect data (e.g., studies with clinical and forensic populations), grant schemes that support multisite work would be particularly useful.

Multisite systematic replication is not yet a guideline for funding research in the social sciences. However, precedents in other fields show how it could be. If clinical-trial phases are considered to be replication attempts, then systematic replication is already quite common in medical research. Also, some medical subfields have funding opportunities for projects that incorporate multisite recruitment. For instance, the National Heart, Lung, and Blood Institute announced such opportunities for clinical-trial applications (National Institutes of Health, Office of Extramural Research, 2018a).

3.2.3. Standardize guidelines for evaluating confirmation-research proposals.

Funding agencies that want to fund confirmation research face two questions: What laboratories and researchers should be funded? And what kind of projects, areas, and target findings should be funded? Regarding the first question, three guidelines are essential. First, proposals should be methodologically sound. Not all replication experiments are created equal, and the literature on replication has provided insights into what constitutes high-quality replication research. For example, high-powered experiments and detailed reporting are elements of good-quality replication work (Brandt et al., 2014). Second, the laboratory needs to meet technical infrastructure requirements (e.g., proposals for developmental psychology studies with infants need facilities that are adequate for infants). Third, the researcher needs to be experienced and to be familiar with the area of study, but this familiarity must not compromise the outcome-independence criterion. In practical terms, this means that replication and original authors should not overlap, and preferably they should not be connected by affiliation or academic genealogy. Regarding the question of what projects to fund, funding agencies should collaborate with the open-science community and professional societies to identify studies that the community regards as worth confirming (see the discussion of interventions by professional societies).

A related precedent in biomedical research is the “rigor and reproducibility” guidance in the grant-application guidelines of the National Institutes of Health (National Institutes of Health, Office of Extramural Research, 2018b). One requirement in this guidance is that the applicant should evaluate the “scientific premise” (i.e., the strengths and weaknesses) of previous research.

Although this requirement (by default) concerns grant applications for novel research, it could be tailored to apply to confirmation research.

3.3. Journals and editors

3.3.1. Publish multisite Registered Reports.

The interventions by universities and funding agencies address the career incentives problem for confirmation researchers. But making multisite replication work sustainable also requires solving the venue problem, which requires the intervention of journals and editors. The key intervention here is opening the door to the publication of multisite Registered Reports and negative results, to give visibility to confirmation researchers' work.

As mentioned earlier, some prestigious journals have begun accepting articles reporting replication studies and negative results. One precedent that also supports the systematicity criterion is this journal's Registered Replication Reports model, which involves multilab work. Still, as of early 2017, only 3% of journals stated that they publish replication research (Martin & Clarke, 2017).

3.3.2. Publish reproducibility reports.

In addition to replication, reproduction is important to confirmation. To encourage such work and reward confirmation researchers further, journals should also publish *reproducibility reports*: short-form peer-reviewed notes presenting the results of running alternative models on the original data, alternative data interpretations, verifications of code, and so on. These reports would be highly beneficial for most subfields, but especially subfields that rely heavily on secondary data analysis. They would also incentivize making maximum use of hard-to-acquire data in, for example, developmental and clinical projects.

Scientists are increasingly interested in performing and consulting reproducibility analyses. Although most of the discussion about the current crisis of confidence has focused on replication, reproduction in many instances is sufficient to disconfirm findings, and it can do so efficiently. If an effect is not robust under an alternative data analysis on the same data set, or if the original publication has errors in reported statistics, it would be highly inefficient to conduct a replication experiment. Nowadays, reports on reproducibility work are for the most part published in blogs. However, journals should both reward high-quality reproducibility work and facilitate it by discouraging original experimenters from managing their data privately. Additionally, if journals link reproducibility reports to the corresponding original publications, then consumers of original articles could access relevant information, and there would be an open record of the evolution of discussions.

3.4. Professional societies

3.4.1. Identify what is worth replicating in each subfield.

In the real world, we cannot estimate every effect size accurately or test every hypothesis. Professional societies should help all the other stakeholders to identify studies that require replication and to assign priority given epistemic and practical payoffs. Here are three general guidelines. First, findings that have already been the target of replication research should receive less weight than other findings (e.g., at this point, we probably do not need more direct replications of classic anchoring effects). Second, findings that have few outcome-independent replications but have inspired a large number of studies should receive more weight than well-replicated and less influential studies, in order to identify potentially misleading bodies of literature. Third, findings that inform (or could inform) evidence-based decisions and treatments (e.g., in clinical and forensic psychology) may require closer scrutiny. Professional societies could collaborate with practitioners to identify such findings.

Identifying replication targets also constitutes a good opportunity for crowdsourced science. Some initiatives already help the community keep track of confirmation and disconfirmation of findings (e.g., curatescience.org). Also, predictive markets can be used to obtain implicit knowledge from the community about what needs to be replicated (Dreber et al., 2015).

3.4.2. Define subfield-specific confirmation guidelines.

As mentioned before, proposals to increase replicability do not apply automatically to all research domains (Tackett et al., 2017). In particular, there is need to adapt the general open-science and good-practices recommendations to different subfields. This intervention is necessary regardless of the self-corrective scheme implemented. However, in the professional scheme, this intervention would help universities and departments to formulate subfield-specific professional profiles (i.e., sets of required knowledge and skills) for confirmation researchers.

Although most publications about increasing replicability have focused on social and cognitive psychology, some authors have already characterized challenges and proposed guidelines in other subfields. Here are two examples. First, consider the generic advice to "get more participants" to increase statistical power. Even though it is sensible, this advice needs to be tailored for specific subfields because sampling procedures can vary significantly from one subfield to another. For instance, in forensic and clinical psychology, projects involving populations with low base rates (e.g., inmates or patients with rare mental conditions) have serious limitations. Specific guidelines to increase power in such contexts may include measurement harmonization and cross-site collaboration to pool data (Tackett et al., 2017). Second, the advice to "directly replicate your own work" might not be constructive in the case of longitudinal studies in developmental psychology that take decades. Instead, a more realistic and useful way to increase replicability is to conduct within-study robustness checks, for example, using multiple estimation techniques and secondary data analysis (Duncan, Engel, Claessens, & Dowsett, 2014).

3.5. Further implementation worries

Before concluding, I discuss some potential remaining concerns about implementing the professional scheme.

3.5.1. Who would like to do only "uninteresting" replication work?

We may worry about whether part of the scientific community would want to do only confirmation research. This is a legitimate concern given that many scientists still deem replication as uninteresting work for "second stringers." (To be fair, however, different subfields in psychology differ in their assessments of the relative value of novel and confirmatory research. For instance, cross-cultural psychology puts confirmation at the forefront given its fundamental interest in population generalizability.) But the categories of "interesting research" and "epistemically valuable research" often do not overlap. Indeed, the more science progresses, the harder it might be to find their intersection. If we cannot find it, we cannot let ourselves be carried away by our appetite for interesting stories and expect an invisible hand to clean up the mess.

From a practical perspective, however, what counts as interesting largely depends on the incentives. Although the thrill of finding something new motivates many scientists, it would be incorrect to think that such a thrill is necessary to motivate them. Many scientists enjoy conducting complex experiments more than developing theoretical innovations. Additionally, academic research is increasingly competitive. Hence, if a job in confirmation research were a career option, it seems that some highly qualified Ph.D. students and postdoctoral researchers would want to do this kind of research.

Finally, consider that society in many contexts relies on workers who do "uninteresting" work that we value. Software testers and food-safety auditors are examples of such workers. Because society acknowledges the importance of their jobs, there are institutions in place to reward them. Confirmation work in science should not be different. The more you think that confirmation efforts are necessary but uninteresting, the more you should agree that we should create an independent system that rewards them.

3.5.2. Would the professional scheme create scientific classism?

One could also worry that separating positions for novel and confirmation research could create classism, putting confirmation researchers in a lower and unappreciated tier. (Think of the worrisome U.S. adjunct faculty system, in which teachers at many institutions are underpaid and overworked relative to permanent faculty.) This is a serious concern given that many researchers (at least in psychology) have not fully acknowledged the importance of replication. The key to preventing such classism is to conceive novel and confirmation research as different in kind but not in value. In practice, this means that universities should design confirmation-research positions taking into account at least two preventive guidelines. First, confirmation-research

tracks should offer the same career-development paths that are available to discovery researchers (e.g., tenure and permanent contracts). Second, universities should not create salary and compensation hierarchies that put confirmation researchers below discovery researchers.

3.5.3. Would the professional scheme affect the incentives for novel research negatively?

Assume that you are doing novel research and you worry about being publicly questioned by independent confirmation researchers. This situation incentivizes two opposed strategies. First, you might want to make your research methodology easily replicable so that independent replications would likely succeed. This would be generally good. You might also decide to play it safe and pursue less creative and less risky (and arguably more replicable) projects than you otherwise might. This possibility, which could seem worrisome at first glance, is not necessarily bad. Indeed, it would be better for the community to rely on modest replicable findings than on creative nonreplicable findings.

The second strategy would be to make your designs extremely complex so that no confirmation researcher would choose to replicate them. This is certainly unwanted. Nonetheless, this strategy would yield only a small payoff in the professional scheme, as you would not contribute to the replication culture. Highly creative research that facilitates replication attempts (e.g., protocols are explicit and materials are shared openly) would be acknowledged the most and therefore incentivized. Producing extremely complex research to steer confirmation researchers away would ultimately harm you.

Finally, these possible strategies rely on the assumption that producers of original research worry about being questioned. However, if independent confirmatory practices become the norm, they will reveal that replication failure should be routinely expected. Hence, scientists would not perceive the possibility that other researchers will fail to replicate their findings as negatively as they do today.

3.5.4. Would the professional scheme slow down scientific progress?

Implementing the professional scheme would require allocating part of the research budget consistently and exclusively to confirmatory research. This change, of course, would not come without cost: The budget that the current system allocates to high-risk exploratory research would be cut. And given this cut, one might worry that scientific progress would slow down. But this worry arises from a mistaken assessment of our current rate of progress. Scientific progress requires reliance on trustworthy findings (i.e., findings that we are willing to accept with a tolerable level of risk). But given the studies suggesting that, in many fields, many results are not replicated, we can reasonably doubt that the current published record is in general trustworthy. Hence, reliance on dubious results could be leading us to misleading bodies of literature and illusory progress. (And this is not merely a theoretical possibility arising from skepticism.) Under the professional scheme, science might *seem* slower, but it would be safer.

3.5.5. Would the professional scheme be cost-efficient?

An objector could worry that the professional scheme may not be a cost-efficient way of addressing the current issues regarding self-correction in science. I have three responses. First, if we trust the low estimates of published findings' rate of replicability (e.g., Open Science Collaboration, 2015), then we have to acknowledge that the status quo is very inefficient: A majority of resources are wasted on findings that turn out to be nonreplicable and uninformative. Hence, we need to invest resources in interventions to reverse this situation.

Second, the professional scheme does not necessarily require large amounts of resources. Relative to discovery researchers, confirmation researchers should be a small group in the scientific community. We do not need to devote (and we should not devote) resources to confirm all findings; rather, we should apply resources so as to maximize information gain for the community. Additionally, the number of scientists who acknowledge the importance of improving their experimental practices (e.g., by not *p*-hacking) is increasing. Hence, looking forward, the work of discovery researchers should be less prone to error. Thus, the number of confirmation researchers needed to keep errors to a reasonable rate should be comparatively small, and most resources would still be invested in novel research.

Third, although other schemes are less resource intensive than the professional scheme, without further adjustments, they are not clearly more cost-efficient, as they have shortcomings meeting the three evaluation criteria. Adjusting those schemes to better meet the criteria would very likely involve incorporating interventions to create distinct rewards for confirmation work; such interventions would be costly and also pretty much in line with the professional scheme. The strongest case can be made for such an adjusted version of the multisite scheme, which would resemble the version of the professional scheme with hybrid novel-research/confirmation-research positions.

A more precise economic evaluation of the professional scheme would require quantitative analysis. This analysis could be done via agent-based simulations to explore cost-benefit trade-offs. For instance, under several assumptions about distributions of agent types (e.g., discovery researchers, replication researchers) and experiment costs (e.g., sample sizes), one could explore how to achieve a specific epistemic goal, such as increasing the replicability rate from 39% (Open Science Collaboration, 2015) to 60% efficiently. I leave this analysis for future work.

3.5.6. Would the professional scheme hinder theory development?

Theory development requires feedback between novel research and confirmation efforts. And one could worry that separating discovery and confirmation researchers could lead each group to produce research that is inconsequential to the other (e.g., confirmation researchers may produce highly reliable research that does not inform new theories, and discovery researchers may produce very innovative but hard-to-confirm theories). However, it is worth noticing that in the professional scheme, researchers doing novel work would still be expected to produce reliable

work to the extent that they can. On the other hand, the intervention of professional societies is key. Regardless of the scheme, professional societies should help their members to keep track of the confirmation status of theories and should produce continuous recommendations about what is important to confirm. This would help confirmation researchers to focus on useful replication efforts.

3.5.7. Given that novel research is increasingly incentivized to be reliable, will the professional scheme cease to be necessary?

These days, the work of confirmation researchers is perhaps more necessary than it has ever been. At the same time, it is also true that most scientists are changing their practices and have higher incentives to produce replicable research. Hence, one could speculate that if the professional scheme were implemented now, confirmation researchers down the road might have less work to do as a result of most novel research becoming reliable. In such a scenario, the professional scheme might no longer be necessary. On the contrary, however, even in the best-case scenario (i.e., that the current crisis will lead to sustainable changes on the side of discovery researchers), there will always be a need for independent and systematic confirmation efforts: Eliminating researcher degrees of freedom entirely is impossible, so there will always be the possibility of mistakes and bias that will need to be corrected. Independent and systematic confirmation efforts should be adopted as standard practice.

Conclusion

Perhaps for scientists in the future it will be evident that any thriving scientific community requires a dedicated group of researchers to keep the community's errors in check. Today the need for such a group is not evident. In this article, I have proposed supporting such a group as part of a self-corrective-labor scheme. The argument consisted of three steps. First, I argued that replication labor (and confirmation efforts more broadly) should satisfy three criteria, namely, systematicity, outcome independence, and sustainability. Second, I used these criteria to evaluate four prominent self-corrective-labor schemes and argued that they all fall short. And third, I argued that the key to overcoming their limitations is a particular division of cognitive labor: A scheme that separates researchers (and their rewards systems) so that some do primarily novel research and others do confirmatory research satisfies the criteria better than schemes lacking this division. To approach this theoretical ideal, I have proposed creating and sustaining confirmation-research-track positions by aligning interventions at the levels of different stakeholders (from funding agencies to journals). Researchers in these positions would be able to conduct replication work that the community needs without the conflicting pressures of our current novelty-based reward system. Although these interventions might seem taxing and detrimental to novel research, they would make scientific progress safer and more certain.

Action Editor

Jennifer L. Tackett served as action editor for this article.

Author Contributions

F. Romero is the sole author of this article and is responsible for its content.

Declaration of Conflicting Interests

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

Funding

This work was partially supported by European Research Council Grant No. 640638, "Making Scientific Inferences More Objective."

Notes

1. Consider some examples. As Yong (2012) documented, when facing the failed attempts to replicate his elderly-priming effect, John Bargh wrote in a now-deleted blog post that the replication researchers (Doyen et al., 2012) were "incompetent or ill-informed" and that there was "nothing in their heads" (Yong, 2012, para. 22). A similar situation occurred after the work of Schnall, Benton, and Harvey (2008), which defended a link between cleanliness primes and moral judgments, was called into question by Johnson, Cheung, and Donnellan (2013). As documented by Meyer and Chabris (2014), advocates of Schnall et al.'s work called the replication researchers "shameless little bullies" and "self-righteous, self-appointed sheriffs" (para. 8; about this case, see also Bohannon, 2014). Another similar controversy concerned the power-pose effect (Carney, Cuddy, & Yap, 2010) and a large-sample failure to replicate it (Ranehill et al., 2015; see Aschwanden, 2016, for discussion). Interestingly, Fetterman and Sassenberg (2015) suggested that researchers overestimate the extent to which failed replications damage original researchers' reputations.

2. Assuming equal sample sizes, an original experiment and a replication experiment will yield confidence intervals of approximately the same size. The difference is negligible given that the width of a confidence interval is primarily dependent on the sample size.

3. A multistudy article presents a series of experiments (mostly conceptual replications) purported to confirm the same general hypothesis. One fundamental problem with these publications is that many of them report only those attempted conceptual replications that yielded statistically significant results, which introduces biases, even in the absence of *p*-hacking.

4. Everett and Earp (2015) suggested a venue such as *PLOS ONE*, which publishes studies regardless of novelty and bases acceptance primarily on whether experiments were well conducted.

5. In other fields, multisite collaboration for replication has become common (e.g., consortia that conduct genomewide-association studies) as a result of editorial enforcement and increasingly

large sample-size requirements intended to ensure that studies are well powered (Kraft et al., 2009).

6. Historians of science regard Bacon's ideas as a major influence in the rise of institutions (e.g., the Royal Society of London) to organize and promote scientific work in the mid-1600s. However, there is a difference in emphasis in how modern science divides cognitive labor and how Bacon conceived his utopia. From an institutional perspective, division of labor in modern science occurs more between fields than between different steps in a project.

7. In philosophy of science, the distinction has been introduced to discuss scientific research that is not strictly driven by specific hypotheses or theories (Steinle, 1997). In recent discussions in psychology, the distinction has been used to stress the importance of producing replicable research (Baumeister, 2016; Sakaluk, 2016). In practice, however, the distinction is overlooked by many scientists, as they often report exploratory research as confirmatory.

8. For example, in 2017, the Religious Replication Project at the University of Amsterdam advertised a position for a Ph.D. student that involved "1. conducting a pre-registered replication study; 2. setting up multi-lab pre-registered replication studies and 3. seeking adversarial collaboration with the original authors of theoretically disputed effects within the field of the psychology of religion" ("PhD Position in Psychology of Religion," 2017, para. 2). A similar position also advertised in 2017, at KU Leuven, required research on "robustness and reproducibility of the results of existing data analysis" and "performing replication studies" ("Assistant Research & Education," 2017, para. 2).

References

Aschwanden, C. (2016, March 24). Failure is moving science forward. *FiveThirtyEight*. Retrieved from https://fivethirtyeight.com/features/failure-is-moving-science-forward/

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27,* 108–119. doi:10.1002/per.1919

Assistant research & education in quantitative psychology II. (2017). *AcademicTransfer*. Retrieved from https://www.academictransfer.com/40402

Bacon, F. (2000). *New Atlantis*. Retrieved from www.gutenberg.org/ebooks/2434 (Original work published 1627)

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7,* 543–554. doi:10.1177/1745691612459060

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71*, 230–244. doi:10.1037/0022-3514.71.2.230

Baumeister, R. F. (2016). Charting the future of social psychology on stormy seas: Winners, losers, and recommendations. *Journal of Experimental Social Psychology, 66*, 153–158. doi:10.1016/j.jesp.2016.02.003

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*, 407–425. doi:10.1037/a0021524

Bissell, M. (2013). Reproducibility: The risks of the replication drive. *Nature, 503*, 333–334. doi:10.1038/503333a

Bohannon, J. (2014). Replication effort provokes praise—and 'bullying' charges. *Science, 344*, 788–789. doi:10.1126/science.344.6186.788

Bones, A. 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*, 307–309. doi:10.1177/1745691612441216

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, England: John Wiley & Sons.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., . . . van 't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217–224. doi:10.1016/j.jesp.2013.10.005

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 365–376.

Campbell, K. E., & Jackson, T. T. (1979). The role and need for replication research in social psychology. *Replications in Social Psychology, 1*(1), 3–14.

Carney, D. R., Cuddy, A. J., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science, 21*, 1363–1368.

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science, 9*, 40–48. doi:10.1177/1745691613513470

Chambers, C. D. (2013). *Registered Reports*: A new publishing initiative at *Cortex*. *Cortex, 49*, 609–610. doi:10.1016/j.cortex.2012.12.016

Chen, R., Desai, N. R., Ross, J. S., Zhang, W., Chau, K. H., Wayda, B., . . . Krumholz, H. M. (2016). Publication and reporting of clinical trial results: Cross sectional analysis across academic medical centers. *BMJ, 352*, i637. doi:10.1136/bmj.i637

Cooper, M. L. (2016). Editorial. *Journal of Personality and Social Psychology, 110*, 431–434. doi:10.1037/pspp0000033

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

Dijksterhuis, A. (2013, April 25). Replication crisis or crisis in replication? A reinterpretation of Shanks et al. [Comment on Article e56515]. Retrieved from http://www.plosone.org/annotation/listThread.action?root=64751

Dijstelbloem, H., Huisman, F., Miedema, F., & Mijnhardt, W. (2013). *Why science does not work as it should and what to do about it*. Retrieved from http://www.scienceintransition.nl/app/uploads/2013/10/Science-in-Transition-Position-Paper-final.pdf

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLOS ONE, 7*(1), Article e29081. doi:10.1371/journal.pone.0029081

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., . . . Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences, USA, 112*, 15343–15347. doi:10.1073/pnas.1516179112

Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2014). Replication and robustness in developmental research. *Developmental Psychology, 50*, 2417–2425. doi:10.1037/a0037996

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., . . . Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology, 67*, 68–82. doi:10.1016/j.jesp.2015.10.012

Everett, J. A. C., & Earp, B. D. (2015). A tragedy of the (academic) commons: Interpreting the replication crisis in psychology as a social dilemma for early-career researchers. *Frontiers in Psychology, 6*, Article 1152. doi:10.3389/fpsyg.2015.01152

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLOS ONE, 4*(5), Article e5738. doi:10.1371/journal.pone.0005738

Fetterman, A. K., & Sassenberg, K. (2015). The reputational consequences of failed replications and wrongness admission among scientists. *PLOS ONE, 10*(12), Article e0143723 . doi:10.1371/journal.pone.0143723

Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science, 7*, 600–604. doi:10.1177/1745691612460686

Freely associating. (1999). *Nature Genetics, 22*, 1–2. doi:10.1038/8702

Grahe, J. E., Brandt, M. J., Wagge, J. R., Legate, N., Wiggins, B. J., Christopherson, C. D., . . . LePine, S. (2018). *Collaborative Replications and Education Project (CREP)*. Retrieved from https://osf.io/wfc6u/

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*, 640–648. doi:10.1097/EDE.0b013e31818131e7

Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences, 18*, 235–241. doi:10.1016/j.tics.2014.02.010

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin, 21*, 1161–1166. doi:10.1177/01461672952111004

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 524–532. doi:10.1177/0956797611430953

John Templeton Foundation. (2017). *The Religious Replication Project: Using pre-registered replications and Bayesian statistics to improve the experimental study of religion*. Retrieved from https://www.templeton.org/grant/the-religious-replication-project-using-pre-registered-replications-and-bayesian-statistics-to-improve-the-experimental-study-of-religion

Johnson, D., Cheung, F., & Donnellan, B. (2013). Cleanliness primes do not influence moral judgment. *PsychFileDrawer*. Retrieved from http://psychfiledrawer.org/replication.php?attempt=MTcy

Kerr, N. L. (1998). Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*, 196–217. doi:10.1207/s15327957pspr0203_4

Kitcher, P. (1990). The division of cognitive labor. *Journal of Philosophy, 87*, 5–22. doi:10.2307/2026796

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahnik, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology, 45*, 142–152. doi:10.1027/1864-9335/a000178

Koole, S. L., & Lakens, D. (2012). Rewarding replications. *Perspectives on Psychological Science, 7*, 608–614. doi:10.1177/1745691612462586

Kraft, P., Zeggini, E., & Ioannidis, J. P. A. (2009). Replication in genome-wide association studies. *Statistical Science, 24*, 561–573. doi:10.1214/09-STS290

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Loevinger, J. (1968). The 'information explosion.' *American Psychologist, 23*, 455. doi:10.1037/h0020800

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7*, 537–542. doi:10.1177/1745691612460688

Martin, G. N., & Clarke, R. M. (2017). Are psychology journals anti-replication? A snapshot of editorial practices. *Frontiers in Psychology, 8*, Article 523. doi:10.3389/fpsyg.2017.00523

Merton, R. K. (1957). Priorities in scientific discovery: A chapter in the sociology of science. *American Sociological Review, 22*, 635–659. doi:10.2307/2089193

Meyer, M. N., & Chabris, C. (2014, July 31). Why psychologists' food fight matters. *Slate*. Retrieved from http://www.slate.com/articles/health_and_science/science/2014/07/replication_controversy_in_psychology_bullying_file_drawer_effect_blog_posts.html

National Institutes of Health, Office of Extramural Research. (2018a). *NHLBI policy regarding submission of clinical trial applications*. Retrieved from https://grants.nih.gov/grants/guide/notice-files/NOT-HL-18-611.html

National Institutes of Health, Office of Extramural Research. (2018b). *Enhancing reproducibility through rigor and transparency*. Retrieved from https://grants.nih.gov/reproducibility/index.htm

Netherlands Organisation for Scientific Research. (2016, July 16). NWO makes 3 million available for Replication Studies pilot. Retrieved from https://www.nwo.nl/en/news-and-events/news/2016/nwo-makes-3-million-available-for-replication-studies-pilot.html

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*, 615–631. doi:10.1177/1745691612459058

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods, 48*, 1205–1226. doi:10.3758/s13428-015-0664-2

Nuijten, M. B., van Assen, M. A. L. M., Veldkamp, C. L. S., & Wicherts, J. M. (2015). The replication paradox: Combining studies can decrease accuracy of effect size estimates. *Review of General Psychology, 19*, 172–182. doi:10.1037/gpr0000034

Nuzzo, R. (2015). How scientists fool themselves – and how they can stop. *Nature, 526*, 182–185. doi:10.1038/526182a

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*, 657–660. doi:10.1177/1745691612462588

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, aac4716. doi:1126/science.aac4716

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*, 531–536. doi:10.1177/1745691612463401

Pashler, H., Harris, C. R., & Coburn, N. (2011). Elderly-related words prime slow walking. *PsychFileDrawer*. Retrieved from http://www.PsychFileDrawer.org/replication.php?attempt=MTU%3D

Peirce, C. S. (1958). The logic of drawing history from ancient documents. In A. W. Burks (Ed.), *The collected papers of Charles Sanders Peirce* (Vol. IV, pp. 89–107). Cambridge, MA: Belknap Press. (Original work published 1901)

PhD position in Psychology of Religion. (2017). *psychoneuroxy.com.* Retrieved from http://www.psychoneuroxy.com/announcement,a2886.html

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., & Weber, R. A. (2015). Assessing the robustness of power posing: No effect on hormones and risk tolerance in a large sample of men and women. *Psychological Science, 26*, 653–656. doi:10.1177/0956797614553946

Reichenbach, H. (1938). *Experience and prediction*. Chicago, IL: University of Chicago Press.

Reproducibility Initiative receives $1.3M grant to validate 50 landmark cancer studies [Press release]. (2013, October 16). Retrieved from https://cos.io/about/news/reproducibility-initiative-receives-grant-validate-50-landmark-cancer-studies/

Roediger, H. L., III. (2012, February). Psychology's woes and a partial cure: The value of replication. *Observer, 25*(2), 9, 27–29.

Romero, F. (2016). Can the behavioral sciences self-correct? A social epistemic study. *Studies in History and Philosophy of Science Part A, 60*, 55–69. doi:10.1016/j.shpsa.2016.10.002

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review, 18,* 682–689. doi:10.3758/s13423-011-0088-7

Sakaluk, J. K. (2016). Exploring small, confirming big: An alternative system to the new statistics for advancing cumulative and replicable psychological research. *Journal of Experimental Social Psychology, 66,* 47–54. doi:10.1016/j.jesp.2015.09.013

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129. doi:10.1037/1082-989X.1.2.115

Schnall, S., Benton, J., & Harvey, S. (2008). With a clean conscience: Cleanliness reduces the severity of moral judgments. *Psychological Science, 19,* 1219–1222. doi:10.1111/j.1467-9280.2008.02227.x

Schönbrodt, F., Heene, M., Maier, M., & Zehetleitner, M. (2015). *The replication-/credibility-crisis in psychology: Consequences at LMU?* Retrieved from https://osf.io/nptd9/

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. doi:10.1177/0956797611417632

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to Registered Replication Reports at *Perspectives on Psychological Science*. *Perspectives on Psychological Science, 9,* 552–555. doi:10.1177/1745691614543974

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26,* 559–569. doi:10.1177/0956797614567341

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *P*-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143,* 534–547. doi:10.1037/a0033242

Smith, N. C. (1970). Replication studies: A neglected aspect of psychological research. *American Psychologist, 25,* 970–975. doi:10.1037/h0029774

Spellman, B. A. (2015). A short (personal) future history of Revolution 2.0. *Perspectives on Psychological Science, 10,* 886–899. doi:10.1177/1745691615609918

Standing, L. G., Grenier, M., Lane, E. A., Roberts, M. S., & Sykes, S. J. (2014). Using replication projects in teaching research methods. *Psychology Teaching Review, 20*(1), 96–104.

Steinle, F. (1997). Entering new fields: Exploratory uses of experimentation. *Philosophy of Science, 64,* S65–S74.

Stephan, P. E. (2012). *How economics shapes science*. Cambridge, MA: Harvard University Press.

Strevens, M. (2003). The role of the priority rule in science. *Journal of Philosophy, 100*, 55–79. doi:10.5840/jphil2003100224

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., . . . Shrout, P. E. (2017). It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science, 12*, 742–756. doi:10.1177/1745691617690042

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*, 1–2. doi:10.1080/01973533.2015.1012991

van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Conducting meta-analyses based on *p* values: Reservations and recommendations for applying *p*-uniform and *p*-curve. *Perspectives on Psychological Science, 11*, 713–729. doi:10.1177/1745691616650874

van Assen, M. A. L. M., van Aert, R. C. M., Nuijten, M. B., & Wicherts, J. M. (2014). Why publishing everything is more effective than selective publishing of statistically significant results. *PLOS ONE, 9*(1), Article e84896. doi:10.1371/journal.pone.0084896

Vazire, S. (2015). Editorial. *Social Psychological & Personality Science,7*, 3–7. doi:10.1177/1948550615603955

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100*, 426–432. doi:10.1037/a0022790

Weisberg, M., & Muldoon, R. (2009). Epistemic landscapes and the division of cognitive labor. *Philosophy of Science, 76*, 225–252. doi:10.1086/644786

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid *p*-hacking. *Frontiers in Psychology, 7*, Article 1832. doi:10.3389/fpsyg.2016.01832

Wright, A. G. C. (2017). The current state and future of factor analysis in personality disorder research. *Personality Disorders: Theory, Research, and Treatment, 8*, 14–25. doi:10.1037/per0000216

Yong, E. (2012, March 10). A failed replication attempt draws a scathing personal attack from a psychology professor [Web log post]. *Discover Magazine Blog*. Retrieved from http://blogs.discovermagazine.com/notrocketscience/2012/03/10/failed-replication-bargh-psychology-study-doyen/

Yong, E. (2013). Psychologists strike a blow for reproducibility: Thirty-six labs collaborate to check 13 earlier findings. *Nature*. Retrieved from https://www.nature.com/news/psychologists-strike-a-blow-for-reproducibility-1.14232