Understanding the Replication Crisis as a Base Rate Fallacy

Alexander Bird

Abstract

The replication (replicability, reproducibility) crisis in social psychology and clinical medicine arises from the fact that many apparently well-confirmed experimental results are subsequently overturned by studies that aim to replicate the original study. The culprit is widely held to be poor science: questionable research practices, failure to publish negative results, bad incentives, and even fraud. In this paper I argue that the high rate of failed replications is consistent with high quality science. We would expect this outcome if the field of science in question produces a high proportion of false hypotheses prior to testing. If most of the hypotheses under test are false, then there will be many false hypotheses that are apparently supported by the outcomes of well conducted experiments and null-hypothesis significance tests with a type-I error rate (α) of 5%. Failure to recognize this is to commit the fallacy of ignoring the base rate. I argue that this is a plausible diagnosis of the replication crisis and examine what lessons we thereby learn for the future conduct of science.

1 Introduction

Psychology and clinical medical science are alleged to suffer from a crisis. In many cases replications of scientific studies purporting to show the existence of certain effects have failed to find those effects or have found effects only of rather smaller size. Many scientists consequently believe that these sciences are suffering from a crisis, and point to problems such as 'questionable research practices' and publication bias to explain how this crisis arose. This paper provides a different explanation for how the replication (replicability, reproducibility) crisis came to be. I argue that the failure of replication should not be regarded as unexpected. Rather the surprise at the number of failed replications and consequent sense of crisis are a result of the fallacy of neglecting the base rate.

In some sciences a high rate of false hypotheses is to be expected. It is not easy to think up correct hypotheses, particularly in fields such as biomedicine. Basic research may suggest links between possible interventions and clinical outcomes. But our incomplete knowledge of the complex systems involved means that one cannot have a high confidence in the correctness of a hypothetical link in advance of clinical trials. So, even if the experimental science that tests the hypotheses is high quality, the high base rate of falsity will show up as a significant proportion of false positives among all the positive results. The replication crisis is complex and polygenic—there are several factors the contribute to it independently. Although the ex-

planation I give is not the only cause of the crisis, it is a significant one. And it is important that it is recognized. First, because it is harmful that a failed replication should immediately be regarded as casting aspersions on the competence and probity of the original scientists. And second, because it must be understood that even if the other factors, such as poor research practices are addressed and eliminated, the problem of irreproducible research may thereby be reduced but it will not go away.

In the light of this result, three approaches to the 'crisis' suggest themselves: (1) Accept that this is the nature of these sciences at their current stage of development, and adjust one's credence in published results accordingly. (2) Seek means of generating research hypotheses that are more likely to be true. (3) Require that the experimental research be of even higher quality, in particular by requiring a rather lower α than the 0.05 currently regarded as acceptable.

1.1 The replication crisis

In certain fields of science, principally social psychology and clinical biomedicine, results that seemed to show certain effects are frequently overturned by subsequent experiments and tests that seek to replicate the original outcome. Often effects that appeared to be firmly established are just not there or are rather weaker than first supposed. The term 'replication crisis' (or 'replicability crisis' or 'reproducibility crisis') is used to describe both the fact that there is this high level of failure to replicate and the sense that these fields face a crisis as a consequence. 52% of 1,576 scientists taking a survey conducted by the journal *Nature* agreed that there is a significant crisis of reproducibility (Baker [2016]).[1] One social psychologist expresses his concerns thus: 'Our problems are not small and they will not be remedied by small fixes. Our problems are systemic and they are at the core of how we conduct our science'.[2]

One reason for anxiety about the crisis, or about talk of a crisis, is that it might accelerate the decline in public trust in science. While trust in scientists remains high in the UK, it has

[1]This article in the *Atlantic* entitled 'Psychology's Replication Crisis Can't Be Wished Away' gives a good sense of the angst being generated and how it is reported in public: <www.theatlantic.com/science/archive/2016/03/psychologys-replication-crisis-cant-be-wished-away/472272/>.

[2]Michael Inzlicht 'Reckoning with the Past', available at <michaelinzlicht.com/getting-better/2016/2/29/reckoning-with-the-past>.

fallen, according to a poll carried out in 2015, by 4% in the preceding year.[3] In the U.S. trust in scientists is lower than trust in the military (Funk and Kennedy [2016]). A particular worry is that the crisis of replication may fuel distrust in science that is far removed from those areas in which the crisis arises. For example, an article in *Investor's Business Daily* entitled 'Memo to Global Warming Alarmists: Science Is Often Wrong' refers to an article that discusses replication failures.[4]

Begley and Ellis ([2012]) report that scientists from Amgen, the biotech company, tried to replicate fifty-three studies in oncology and related fields over a ten-year period. Of these only six (11%) were satisfactorily replicated. Another study (Ioannidis [2005a]) considered all the most highly cited clinical research studies published in three leading medical journals, and looked to see whether there were subsequent studies of the same hypotheses that were larger or better controlled. 45 such studies claimed an effective intervention. 16% of these were contradicted by the later studies, while in another 16% the effect was found to be smaller in the later than in the earlier study. 44% were replicated and the results of the remaining 24% of studies went largely unchallenged.

A number of very significant results in psychology have failed attempts at replication. One very well-known and influential result in social psychology (Bateson *et al.* [2006]) says that an image of a pair of eyes watching a subject will tend to increase their socially co-operative behaviour, such as the amount the subject pays into an honesty box for tea and coffee. But a larger study (Carbon and Hesslinger [2011]) failed to find the same effect. Another commonplace of social psychology is that babies are born imitators. Yet this widely accepted claim has also been subjected to serious doubt by a high-quality study (Oostenbroek *et al.* [2016]), which found that babies were no more likely to stick out their tongues when the researcher was sticking her tongue out than when the researcher was making some other facial display to the baby. Various priming effects, whereby exposure to certain cues can influence attitudes and behaviour, have also been difficult to replicate (Doyen *et al.* [2012]; Klein *et al.* [2014]). One interesting piece of research (Carney *et al.* [2010]) that has received a lot of publicity

[3] </www.ipsos-mori.com/researchpublications/researcharchive/3685/Politicians-are-still-trusted-less-than-estate-agents-journalists-and-bankers.aspx>.

[4] </www.investors.com/politics/commentary/do-not-trust-climate-science-scientists-wrong-all-the-time/>.

maintained that adopting high-power poses causes increased feelings of power and tolerance to risk, alongside changes in biomarkers (raised testosterone and lowered cortisol). A much larger study (Ranehill *et al.* [2015]) replicated the increased self-reported feelings of power, but did not find any behavioural change (for example, regarding risk-taking), nor any change in the biomarkers. Even effects as central to psychology as ego depletion and stereotype threat have come under suspicion as a result of failed replications (Hagger *et al.* [2016]; Flore and Wicherts [2015]).

1.2 The base rate fallacy

The base rate fallacy is liable to occur when making an inference regarding the probability of some particular occurrence of a general phenomenon (for example, whether an individual has a disease). The fallacy arises when the reasoner focuses solely on some salient piece of evidence regarding that occurrence while neglecting the rate at which occurrences of that phenomenon would occur independently of that evidence (the base rate). This leads to erroneous conclusions when the evidence is strongly but imperfectly correlated with the occurrence (for example, when the evidence is some kind of test for phenomenon) and the phenomenon itself is rare. In a well-used example, a profiling tool might scan airline passengers for appearance and behaviour that is indicative of being a terrorist. The test might be a good test in that it is accurate in 95% of cases. That is, of every 100 non-terrorists, it says that 95 are not terrorists and 5 are terrorists, and of every 100 terrorists it says that 95 are terrorists and 5 are not terrorists. So, if a passenger fails the test—it says he is a terrorist, what is the chance that he is a terrorist? Certainly not 95%. The chances are in fact miniscule. For the number of terrorists is tiny compared to the number of ordinary air passengers. So if a passenger fails the test it is far, far more likely that he is an innocent passenger who is a victim of the 5% failure rate of the test then he is a genuine terrorist accurately caught by the test.[5]

The evidence is strong that people are prone to neglect the base rate and so make fallacious

[5]Similar criticisms may be made of other screening programmes for rare outcomes, such as the UK's programme for screening travellers for Ebola (see Robin Evans's comments at </itsastatlife.blogspot.co.uk/2014/10/ebola-hunting.htmlmore/>. As far as I can tell Evans was right in his prediction. In the UK the programme identified 367 people as 'high risk' and a further 19 were referred to the NHS. None of these had Ebola.

inferences, such as inferring that the passenger is likely to be a terrorist. In a famous experiment (Casscells *et al.* [1978]) medical students at Harvard were asked to say how likely it was that a patient (without other indications or risk factors) who tested positive for a rare disease in fact has the disease. The students were told that the disease is found in 1 in 1000 patients and that the test is 95% accurate. Almost half the students said that the probability that the patient has the disease is 95%. In fact, the probability is less than 2%—only 11 of the 60 students got this correct answer.

We can see why the right answer is approximately 2% as follows: Let π be the probability that S has the disease, independently of the test result; and so the probability that S does not have the disease is $1 - \pi$. Let $\mathbf{r}$ be the accuracy of the test; so the probability of the test yielding an inaccurate result is $1 - \mathbf{r}$. See Table 1 for the various probabilities.

Table 1

	Has disease	Has no disease
Tests positive	$\pi\mathbf{r}$	$(1-\pi)(1-\mathbf{r})$
Tests negative	$\pi(1-\mathbf{r})$	$(1-\pi)\mathbf{r}$

(Note that in this example, as in the terrorist detection case, what we have called 'accuracy' in fact covers two kinds of accuracy: the probability that someone with the disease tests positive and the probability that someone without the disease tests negative. For many tests these will in fact be different and I will differentiate them later in the discussion.)

What we want to know is:

$$\text{Pr(S has the disease, given that S tests positive).}$$

This is equal to:

$$\frac{\text{Pr(S has the disease and S tests positive)}}{\text{Pr(S tests positive)}}.$$

The denominator, Pr(S tests positive), is equal to: (Pr(S has the disease and S tests positive) + Pr(S does not have the disease and S tests positive)). So, using Table 1:

$$\text{Pr(S has the disease, given that S tests positive)} = \frac{\pi\mathbf{r}}{\pi\mathbf{r} + (1-\pi)(1-\mathbf{r})}.$$

We can plug in the values given in the Harvard study: $\pi = 0.001$ and $\mathbf{r} = 0.95$. Which gives:

Pr(S has the disease, given that S tests positive) = 0.019 = 1.9%.

This result shows that although the test is highly accurate (95%), the probability that a positive test correctly indicates a case of the disease is very low (<2%). The reason for this is the fact that the disease is very rare, so a positive test is much more likely to arise from a non-diseased case giving a false positive than from a diseased case giving a true positive.

2 From the Base Rate Fallacy to the Replication Crisis

This section connects the two preceding sections by showing how the large number of failed replications can be explained by reference to a high base rate of falsity among the hypotheses tested. The sense of 'crisis' arises, I suggest, from the fallacy of ignoring this base rate. In what follows I assume that the scientists in question are concerned with scientific research hypotheses that are usually causal in form. For example, a research hypothesis may state that a new drug will bring about a reduction in blood pressure for a particular cohort of patients. Or it might state that certain stimuli will cause subjects to recall a certain type of fact more quickly. Scientists have standard means of testing such hypotheses. Such tests are imperfect, in that they can produce false positives (false hypotheses that are accepted as true) and false negatives (true hypotheses that are accepted as false). For example, hypotheses of these kinds will typically be tested with a randomized controlled trial (RCT) whose results are analysed by null hypothesis significance testing (NHST), with the significance level set at 5%. That is, patients or subjects will be randomly allocated to a control group and an experimental group and the difference in outcome is measured. A statistical null hypothesis is formed, stating that there is no baseline difference in outcome between the two groups. Using NHST we calculate the probability that we would see the observed difference in outcome (or a larger one) if the null hypothesis is true. If that probability is less than 5%, the null hypothesis is rejected. Correspondingly, the research hypothesis is accepted—it will be published as a statistically significant outcome. Setting the significance level at 5% means that if the null hypothesis is true, then there is a 5%

probability of falsely rejecting it. This method of hypothesis testing (RCT plus NHST plus a 5% significance level) is only 95% accurate in this respect—it accepts a false positive rate of 5%.[6]

Consider the mad scientist Dr M who has many crazy ideas. He pursues wild hypotheses, only 1 in 1000 of which is true. (He may be influenced by Popper's exhortation to devise hypotheses that are bold.)[7] However, he is a responsible enough scientist to test his hypotheses using accurate methods, properly conducted, to the standards expected by the scientific community. For example, he may employ randomized controlled trials with null hypothesis significance testing, with the significance level set at 5%. Let us assume then that his test method is 95% accurate in declaring that a true hypothesis is true and that false hypothesis is false. Given that a test says that one of Dr M's theories is true, what in fact are the chances of its being true? This case is exactly analogous to the disease case. Instead of a case of a disease we have a true hypothesis, and instead of a test for the disease we have a test for truth. We can conclude that the theory's probability of bring true is only 2%.

Now consider Prof. S. She is entirely sane, but works in a new and difficult area where there is little solid background theory or reliable results to guide her hypothesis formation. Rather she uses her scientist's intuition or analogues with other ideas and results to generate hypotheses that she then submits to careful scrutiny. Let it be that 10% of the hypotheses she forms are true. And as before, assume that her test methods are 95% accurate. We can now draw up the following table:

[6]Some commentators assert that 'the null hypothesis, taken literally, is always false' (Meehl [1978], p. 822; see also Cohen [1990], p. 1308). In which case, we cannot usefully talk of the probability of falsely rejecting it. Cohen's example of a null hypothesis is 'The difference in the mean scores of US men and women on an attitude toward the UN scale is zero'. Regarding such cases where we are comparing two distinct actual groups in a population, Cohen is surely right. Matters are more contentious, however, when the study uses an RCT (Hagen [1997]; Lakens [2014]). And Meehl ([1990], p. 204) himself notes the the equivalent claim that 'everything correlates to some extent with everything else' is not true for pure experimental studies (such as an RCT). In any case, for our purposes, we need only note two things. First, that the negations of the scientific research hypotheses, the 'causal null hypotheses', so to speak, that assert that there is no causation between one factor and another, are not always false and are very frequently true. And second, that using this method to test them will generate a false rejection of the scientific null and so an incorrect acceptance of the positive research hypothesis in 5% of cases where the research hypothesis is false (that is, when the causal null hypothesis is true).

[7]'Bold ideas, unjustified anticipations, and speculative thought, are our only means for interpreting nature' (Popper [1959], p.280). If scientists generate ideas with this in mind one might expect a very high background rate of falsity among new hypotheses, as Popper himself held.

Table 2

	Hypothesis is true	Hypothesis is false
Passes test	0.1×0.95	0.9×0.05
Fails test	0.1×0.05	0.9×0.95

and calculate the probability that one of Prof. S's hypotheses, chosen at random, is true if the test says it is true—which is 0.68. So only two-thirds of S's hypotheses that are successful in passing the test for truth, are in fact true. That means that one third of her successful theories are in fact false. The key to understanding why is the base rate of falsity in the hypotheses she produces. Since 90% of her hypotheses are false, a large proportion of her successes will in fact be false hypotheses that came up with type-I errors (false positives) in the test. Now let us imagine that other scientists come along and test S's successes. Since a third of her successful hypotheses are false, that falsity will mostly show up in the outcome of these other scientists' tests. Since about a third of S's successful theories are not successful on re-testing, it may look as if S is a shoddy scientist. But S is not a shoddy scientist, in that she carries out high quality experiments that get the correct result 95% of the time. On the other hand, she is disadvantaged by testing hypotheses that are very likely to be false. That fact may have nothing to do with shoddy science and everything to do with the newness or difficulty of the field in question.

While the case of Dr M is fantastical, that of Prof. S is entirely realistic. The key assumption, that only 10% of the hypotheses tested are true is, in fact the case in psychology, according to an analysis by Johnson *et al*. ([2017]).

The central claim of this paper is that the replication crisis can be explained is the same way. Sterne and Davey Smith ([2001]) argue that when hypotheses have a low probability of being true, significance testing can generate a high proportion of positive test outcomes that are in fact false positives, a point noted by Wacholder *et al.* ([2004]) and by Colhoun *et al.* ([2003]) in connection with molecular and genetic epidemiology, and amplified in Ioannidis's ([2005b]) broader claim that 'most published research findings are false'.[8] Just as Prof. S's science produced many false theories that passed her tests, so psychology and biomedicine have

[8]Sterne and Davey Smith ([2001]) also point out the analogy with false positives in screening tests.

produced many false theories that seem to be correct according to the experiments carried out by scientists in those fields. And just as the falsity of Prof. S's many false but successful hypotheses is revealed on re-testing, so in these fields also there is a high level of failure to replicate the experiments supporting successful theories. The remainder of this section articulates this proposition in more detail; the next section provides some reason to think that it might be true.

The first adjustment to make to the base rate fallacy model of the replication crisis is to distinguish between the two types of accuracy discussed above. First, there is accuracy in saying that X is so when X is so. This is described as the power of a study. Second, there is accuracy in saying that X is not so when X is not so. While this does not have a formal name, it is equal to what is often called the confidence level of the study. Correspondingly, there are two kinds of inaccuracy: saying that X is not so when X is in fact so—type-II errors (or false negatives); and saying that X is so, when in fact it isn't so—type-I errors (false positives). The type-II error rate is symbolised by β, and the type-I error rate by α. In summary:

Table 3

Type of inaccuracy	Error rate	Accuracy	Type of accuracy
Type-I error	α	$1-\alpha$	Confidence level
Type-II error	β	$1-\beta$	Power

So we can now display the various probabilities in Table 4:

Table 4

	Hypothesis is true	Hypothesis is false
Passes test	$\pi(1-\beta)$	$(1-\pi)\alpha$
Fails test	$\pi\beta$	$(1-\pi)(1-\alpha)$

The false positive report probability (FPRP) is the probability of a hypothesis being false

even though it passes the test (Wacholder *et al.* [2004]). From Table 4, we see this is:

$$\text{FPRP} = \frac{(1-\pi)\alpha}{(1-\pi)\alpha + \pi(1-\beta)}.$$

We must distinguish

Pr(hypothesis is false, given that it passes the test), which is FPRP

from

Pr(hypothesis passes the test, given that it is false), which is α.

Indeed, mistaking the two, so that one takes a high value for α to equate to a high value for FPRP is a fallacy of probabilistic thinking of which the fallacy of neglecting the base rate is one manifestation. The p-value fallacy is another (Goodman [1999]).

If we set $\alpha = \beta = 0.05$ and $\pi = 0.1$ as in the case of Prof. S, then we get the result, as before, that the probability that hypothesis that passes a test is in fact false, is about one third. As explained above, this result is the outcome of a combination of π being low, plus α, although small, being non-negligible. We can see that as long as the type-II error rate, β, is not close to 1, it does not have much of an effect on FPRP. I will consider the importance of different values of β below.

On the other hand, for FPRP to be small, we need either π to be close to 1 or α to be close to 0. And so if π is itself is not close to 1 but instead is close to 0, then α needs to be very close to zero—that is, negligibly small—for FPRP to be small. For example, let us say that we want it to be the case that when we get a positive test outcome, this is erroneous in only 5% of cases, that is, FPRP = 0.05. As we have seen, it is fallacious to think that setting α to be 0.05 will achieve this. What value should we give then to α to have a FPRP of 5%? If we keep $\pi = 0.1$ and $\beta = 0.05$, then for FPRP to be 1 in 20, the value of α must fall from 0.05 to 0.0056 (one ninth of the value we have been working with). I return to this point in more detail below.

3 Explaining the Replication Crisis: Low π and Non-negligible α

The preceding section argued that if there is both (i) a low background rate of truth among hypotheses proposed and tested and (ii) a significance level set at a value that although small

is not negligible, then we would expect a high proportion of positive results of tests of those hypotheses to be erroneous. Hence we would expect many replication studies to fail to reveal what the original studies seemed to show. It is one thing to argue that low π and non-negligible α would explain the replication crisis, but quite another to show that they actually do. To complete the proposed explanation of the replication crisis, we need to consider whether in fact the relevant sciences are such that they combine low π with non-negligible α.

3.1 Science with low π

This explanation for the replication crisis depends on there being a low rate of true hypotheses among those considered sufficiently seriously to be tested. Is that plausible?

It is difficult to test directly for the hypothesis of low π. A natural way to attempt to do this would be to use repeated tests or tests of higher quality to find out what the proportion of the hypotheses that initially tested positive are in fact true (which is 1 – FPRP). But to do so would be to employ the model I have presented, using the value for FPRP to calculate a value for π. That would be to beg the question as issue, whether this model is a plausible explanation of the replication crisis.[9]

So we have to use indirect considerations. Why might a field have a low π for its hypotheses? Let us start from the other end. Why might a field have a high π? One reason is that the field is dominated by a well-established and well-confirmed theory and the hypotheses of the field test specific aspects of this theory or applications of the theory in particular domains, or they extend the theory in plausible ways. In such cases it is possible to have a fairly high degree of confidence that the hypothesis is true. For example, the discovery of the Higgs boson in 2012 was the experimental verification of a hypothesis that had first been proposed in the 1960s and was developed in subsequent decades. That hypothesis was not a stand-alone theory but was a component part of the standard model of particle physics. The standard model is one of the experimentally best confirmed theories in all science, and before its discovery, the Higgs mechanism was the one remaining unverified part of the model. Furthermore, physicists had

[9]For this reason we cannot use the conclusion of Ioannidis ([2005b]), that most published research findings (in biomedicine) are false, since he uses similar arguments to reach that conclusion. Indeed, Goodman and Greenland ([2007]) accuse Ioannidis of circular reasoning.

a clear understanding, based both in the well-established theory and in a long-standing experimental tradition, of what kind of experiment would detect the Higgs particle if it exists. Of course, despite its past successes, the standard model might be erroneous in important respects and the route from the standard model to the Higgs hypothesis was not trivial and assumptions, albeit plausible ones, need to be made. So the outcome of the experiments carried out using the Large Hadron Collider at CERN in July 2012 were not a forgone conclusion. Nonetheless, most physicists believed that the Higgs particle would be discovered by that research. Elsewhere in physics, the existence of gravitational waves, first directly confirmed in 2016, was a longstanding prediction of Einstein's general theory of relativity. Since the latter has considerable independent experimental support, the gravitational wave hypothesis was highly plausible. Designing and carrying out a suitable detection experiment was not easy. But again, given the track record of success of general relativity, the widespread expectation was that gravitational waves do exist and would produce the relevant outcome in a suitable experiment.

Clinical medicine and psychology, especially social psychology, contrast with physics in two respects. First, although our understanding of physiology and pathology are improving rapidly, the biological systems in question are so complex that our knowledge remains radically incomplete. And the hypotheses under consideration in clinical medicine are not more-or-less direct consequences of those underlying basic theories. Rather they are hypotheses concerning the action of a drug on patients, and so stand at some inferential distance from the underlying theory. So the connection between our underlying theory and the hypothesis under test is much weaker than in physics. Second, the experimental evidence for the underlying theory is generally much weaker also.[10]

Thus in physics, we can say, first, 'if the standard model is correct, then the hypothesis of the Higgs mechanism is very probably true'; and, second, 'there is very strong evidence that the standard model is correct'. Together these imply that it is probable that Higgs hypothesis is correct. Consider, by contrast, the trials of the drug Bapineuzumab, a proposed treatment for Alzheimer's disease. Here the underlying basic theory is that beta-amyloid (Aβ) plaques that

[10]There are other reasons why one science might have a lower π than another, some of which I consider below. One difference, pointed out to me by an anonymous referee, is that some fields have a well-established practice of circulating preprints, working papers, and the like. This may help eliminate errors or implausible hypotheses before formal publication or even before experiments are conducted.

are characteristic of Alzheimer's play a causal role in the pathology of the disease and cause the cognitive impairment that is symptomatic of it. This is the amyloid cascade hypothesis. Bapineuzumab is an antibody to those plaques, and hence it was hoped that the drug would inhibit and reduce the Aβ plaques and so the symptoms of Alzheimer's. That it would do so is the target hypothesis tested in two major trials in 2012. However, in this case, the underlying theory, the amyloid cascade hypothesis, though supported by evidence is not undisputed. Some researchers hold that so-called tangles (collapse of the τ proteins) are an equally or more significant causal factor. The relationship between the Aβ plaques, the tangles, and the cognitive effects of Alzheimer's disease are far from being fully understood. So the theory underlying the hypothesis that Bapineuzumab would be an effective intervention for Alzheimer's is both incomplete and subject to doubt. Hence the best one can say in this case is first, 'if the amyloid cascade hypothesis is correct, it is possible that reduction in plaques will halt or reverse Alzheimer's, and so conceivable that an antibody to the plaques will assist Alzheimer's sufferers'; and, second, 'the amyloid cascade hypothesis may be correct, but the evidence is far from conclusive'. Consequently, the hypothesis that Bapineuzumab would be an effective treatment for Alzheimer's was at best a reasonable hope; it was hardly a hypothesis that anyone should have expected to be true. In summary, mechanistic reasoning is very helpful in medicine, but it often fails to support a high probability (high π) for its hypotheses. Analogous comments may be made about psychology and social psychology where there is little in the way of a complete and well-evidenced underlying theory from which hypotheses can be derived with confidence. In medicine there are additional reasons for a low π. As Begley and Ellis ([2012]) explain: 'Given the high unmet need in oncology, it is understandable that barriers to clinical development may be lower than for other disease areas, and a larger number of drugs with suboptimal preclinical validation will enter oncology trials'. Such a motivation—alongside a financial one—was no doubt at work in trialling Bapineuzumab before the basic science ruled out the tangle hypothesis.

Other areas in non-clinical biomedicine benefit from the fact that it is easy to generate and test new hypotheses. Genetic and molecular epidemiology seeks causal linkages between genetic variants and disease, in particular in connection with environmental factors. As Wacholder

et al. ([2004]) explain, when it was not feasible to study multiple genes in a pathway, let alone an entire genome, studies were directed at particularly promising hypotheses, prompted by strong biological evidence; whereas advances in technology have now made it much easier and cheaper to test for possible relationships between variants in genes (whose function may not be known) and disease. Given the large number of possible relationships, but the rather smaller number of true relationships, it is to be expected that many of the hypotheses tested will be false.[11]

Hypotheses in clinical medicine and in psychology, as well as in other sciences, may come from sources other than a basic theory. Hypotheses may be suggested by the results of observational studies or even by unsystematic observations or the researcher's intuition. These are perfectly legitimate means of generating hypotheses. On the other hand, they are not means of generating hypotheses that give those hypotheses a high probability of truth. Indeed, some sources seem clearly liable to produce false hypotheses. For example, psychologists have been investigating whether subjects who acted out the literal meaning of a metaphor would demonstrate the behavioural dispositions associated with their metaphorical meaning. In one study (Leung *et al.* [2012]), subjects working inside a box-like room were compared with subjects working outside of the room with regard to their creativity to see whether literally 'thinking outside the box' can boost creativity. In another study (Sanna *et al.* [2011]), researchers investigated whether subjects in a physically more elevated position (occupying the 'moral high ground') were more likely to engage in pro-social behaviour (exhibiting 'higher virtues'). It is perfectly legitimate to investigate such ideas. But they are clearly not hypotheses in which one should have had high confidence ahead of testing. Consequently, although the some such experiments produce positive results, it should not be a very great surprise that subsequent attempts at replication do not.

A common motivation or inspiration for a new hypothesis is an analogy with other previously confirmed hypotheses (Dunbar [1996]; Blanchette and Dunbar [2000]).[12] For example,

[11] Ioannidis ([2005b], p. 699) gives an example concerning genetic determinants of schizophrenia.

[12] Thomas Kuhn ([1970]) argued that the key element in his notion of a 'paradigm' is what he called an 'exemplar'—an exemplary solution to a scientific problem that become a model for subsequent research. In his view, normal science is driven by modelling research problems and their solutions on such exemplars.

the power-pose study (Carney *et al.* [2010]) referred to above does not cite any underlying theory from which the hypothesis is derived. Rather the authors refer to two sets of research findings. First, they point out that power is expressed through physical displays, a commonplace observation confirmed by research. Second, they refer to research on embodied cognition:

> [...] some research suggests that bodily movements, such as facial displays, can affect emotional states. For example, unobtrusive contraction of the "smile muscle" (the zygomaticus major) increases enjoyment (Strack, Martin, Stepper, 1988, 88), the head tilting upward induces pride (Stepper and Strack, 1993) etc.

On the basis of these findings the researchers hypothesised that adopting high-power poses would cause increased feelings of power and tolerance to risk. Again, given such background information, it is reasonable to suppose that there might be such an effect. But that information certainly does not make it likely that there is such an effect.

There is a further problem when a false positive result itself becomes an exemplar upon which further research hypotheses are based. Some of the supposed effects that have failed to replicate, such as social priming, are sufficiently accepted and important that new research hypotheses are suggested by analogy with them. With many false positive results, the practice of modelling new hypotheses on analogies and parallels with old results will itself lead to those new hypotheses being false in most cases. And this may well explain the problem with the power-pose hypothesis. For it also turns out that the finding that smiling increases enjoyment, upon which the power-pose hypothesis was explicitly based, has itself failed tests of replication (Wagenmakers *et al.* [2016]). What we may call the 'falsity feedback effect', whereby a study with a false positive becomes an exemplar (model) for subsequent hypotheses, will clearly encourage a higher base rate of falsity among hypotheses.

3.2 The value of α: Type-I errors

For very many studies in the fields we are considering α is set as 0.05, the value we have been using above. That is because what counts as passing the experimental test is that it should produce a p-value of less than 0.05. When an experimental test is carried out, the observed

value of some parameter of interest may differ between the experimental arm of the experiment and the control arm, in a way that is qualitatively in line with the hypothesis. Some such difference may come about by random error, even if the hypothesis is false. When the data is analysed, its *p*-value is calculated. This is the probability of obtaining a difference between the two group at least as great as that observed, if the null hypothesis is true (that is, if the hypothesis under test is false). This is null hypothesis significance testing. In defining success as a *p*-value as less than 0.05, we are saying that the chances of a successful (positive) result given that the hypothesis is false is 0.05. We are setting success at a level such that type-I errors occur in 5% of false hypotheses tested.

It is notable that this threshold is far less stringent than that used in some other sciences, most obviously physics. When the Higgs boson was discovered, the data met the widely accepted 'five sigma' (5σ) standard. Here σ is one standard deviation from the mean. So 5σ refers to an outcome that is five standard deviations (or more) from the expected mean value, should the null hypothesis be true. That is equivalent to an α of 0.00003%, which is to say we would expect to get this positive result (or more extreme) from a false hypothesis on, roughly, one occasion in every 3 million tests. The principal reason for physics using this standard is the quantity of data produced by the experiments in particle physics. This has two consequences. First, it is possible to produce data that meets this standard, whereas in medicine and psychology is just is not feasible to have enough patients or subjects to generate data of that quality. Second, it is necessary to have data that meets this high standard, in order to avoid false positives. This is because scientists use the data from experiments to look for patterns—so-called bumps in the data—that may be suggestive of hitherto unsuspected effects. However, given the large quantity of data from a noisy environment, one can expect this kind of data-trawling to generate false positives—false hypotheses arising from statistical blips. So a correspondingly stringent standard of discovery is required. For example, several bumps were detected by Fermilab's Tevatron collider that were significant at the 3σ level and which subsequently disappeared when more data was acquired. Physicists thus typically regard data at the 3σ level as 'evidence' of an effect (such as the existence of a new particle) while 5σ is required for 'discovery'. In effect, the physicists are acknowledging the claim made in this paper: if you use a low quality means of

generating hypotheses (a means that generates a high background rate of falsity), then you need a correspondingly high standard of statistical significance (that is, a very low α). Additionally, the external pressures are different for physics from medicine. The huge cost of the particle accelerators in physics means that the physicists running them have strong reason to avoid retracting announcements of discoveries. Whereas in medicine there is economic pressure on drug companies to produce results, as well a moral pressure to find cures.

4 Other Explanations of the Crisis

The *Nature* survey reported that 60% of respondents held that pressure to publish and selective reporting played a major part in the crisis surrounding failed replications. More than 50% referred to insufficient replication in the lab, poor oversight, or low statistical power. None pointed to the two factors identified in this paper: low π and too high a value for α.

4.1 Low statistical power

Some commentators have associated the replication crisis with the low statistical power of many studies in the relevant fields. Statistical power is the converse of β, the type-II error rate. It is one of the two kinds of accuracy considered above, $1 - \beta$ (the other being $1 - \alpha$).

Button *et al.* ([2013]) provide an analysis to support this claim. They focus on the positive predictive value (PPV) of a positive outcome in a test or experiment: the probability that a hypothesis is in fact true, given that it yields a positive outcome. This is the number 1 − FPRP I mentioned above. We can see straightforwardly that:

$$\text{PPV} = 1 - \text{FPRP} = \frac{\pi(1-\beta)}{(1-\pi)\alpha + \pi(1-\beta)}.$$

And this number increases as β decreases, and so in order to maximise PPV we should aim to reduce β, the type-II error rate, or, equivalently, increase the power of the study, $1 - \beta$.[13]

I do not find this analysis entirely satisfactory either as an explanation of the replication crisis

[13]The analysis from (Button *et al.* [2013]) is slightly different in that they use the pre-study odds of the hypothesis being true, whereas I have used the π probability of its being true. These are mathematically related so their analysis is mathematically identical to that given in this section.

or as a solution. Note first that the problem does not disappear if we require all the studies to have a very high power. In the example of Prof. S we assumed an accuracy of 95% with regard to both type-I and type-II errors: $\alpha = \beta = 0.05$. That is a very high power and would be unusual in the areas of research we are discussing.[14] It is rather higher than the 80% typically thought to be satisfactory. Nonetheless, Q's PPV is only 68%—almost a third of her positive results are false. With her π of 0.1 and α set at the conventional 0.05, even maximum power (100%) will increase her PPV only by 1% to 69%. For that matter, if her power were lowered to 80%, then her PPV is still 64%. And lowering her power to the rather feeble 50%, lowers her PPV only to 53%. Thus a power change from a low 50% up to an impossibly perfect 100% raises S's PPV from 53% to 68%. Thus a change in power, in this case, makes only a marginal difference to PPV.

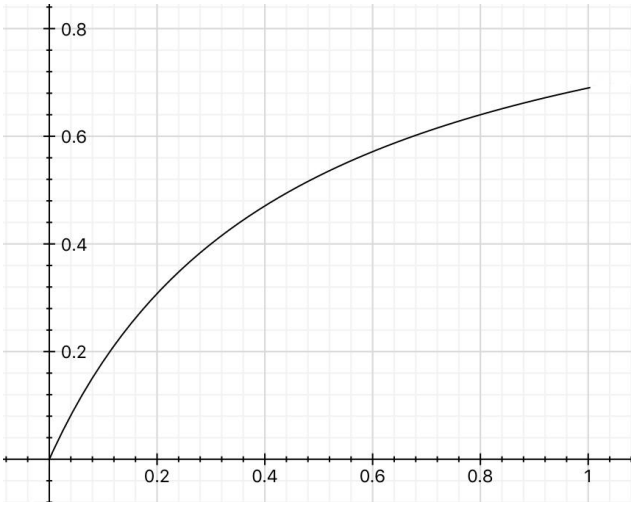

Figure 1: PPV (y-axis) as a function of power (x-axis) when $\pi = 0.1$ and $\alpha = 0.05$). Even with maximum power, almost a third of positive results will be false positives.

The positive experimental results in a field are a combination of the false positives and the true positives. So we have a high chance of a positive result being a false positive when the former are numerous and the latter are rare. It is therefore correct that PPV will improve when we increase the number of true positives. But in a context where there is a high number of false positives, this seems to be focusing in the wrong place. What we really need to be doing is to reduce the false positives. Think of the example of terrorist detection device. In that case what

[14] I have used this unusually high value in order to isolate the contribution of a non-negligible α.

makes the device impractical is the large number of innocent travellers that are falsely identified as terrorists. The overall proportion of terrorists among the positives will be increased if we raise its type-II accuracy from 95% to 100%. But by only a very little and not by enough to render the device useful.

I do note, however, that if power is already rather lower than I have supposed, say around 0.2, then the benefit to PPV of increasing power will be greater, because the gradient of the curve in Figure 1 is steeper for low power. So where in fact do the sciences in question find themselves on the curve in Figure 1? There is recent evidence that in some fields power is low, as low as 0.2 for small effect sizes, in cognitive neuroscience and psychology (Szucs and Ioannidis [2017]) and some subfields of biomedicine (but not others) (Dumas-Mallet *et al.* [2017]). The principal cause of low power is small sample size (small effect size is another important factor). Dumas-Mallet *et al.* ([2017]) excluded studies that concerned treatments of disease (and also screening and diagnosis)—which are studies that would typically have large sample sizes. So for some fields it does appear that low power is a problem, and this does reduce PPV because there are even fewer true positives to counteract the false positives generate by non-negligible α. So increasing power, say from 0.2 to the accepted standard of 0.8, will increase PPV from about one third to two-thirds (for $\alpha = 0.05$ and $\pi = 0.1$). But, as argued above, reducing the remaining third (that is, getting PPV above 0.68) cannot be achieved by any amount of increase in power, but only by reducing α, as Figure 2 shows, or by increasing π.

There are three other respects in which raising power has some benefits, other than by directly increasing PPV. First, and most importantly, we should want the replication tests themselves to be highly powered. Direct replications will often just mirror the original experiment along with its power calculations, which may in any case overestimate the experiment's true power. A low powered replication may fail to reveal a true effect. So in some cases we might mistakenly believe that original hypothesis and its supporting experiment have failed a test of replication, but in fact they are correct and it is the replication that is in error. Note that this error is a type-II error and so power is directly relevant. We should want to increase the power of our replications so that we can be sure that a failure to replicate really is a problem for the original hypothesis and experiment.

Second, when it comes to minimizing the falsity feedback effect it will help to have a higher proportion of the positives being true positives, even if that does not come from reducing the absolute number of false positives. Given limited resources for carrying out future research, it will be better for science if a larger proportion of new hypotheses are modelled on true exemplars, thus increasing π for those hypotheses.

Finally, increasing power will be achieved by increasing the sample size. That in turn means that it will be possible to decrease α while still being able to detect worthwhile effects. That depends on our being willing to reduce α, the choice of which is a matter of convention. If one increases power without reducing α then our experiments will simply reveal more positives where the effect size is small. This may be tempting. But it makes no difference to the analysis just given. It will remain the case that many of the newly detected apparent 'effects' are mere statistical artefacts. I return to this below.

To summarize, I have used an implausibly high power of 95% in order to isolate the effect of α set at 0.05, when $\pi = 0.1$. Even power of an impossible 100% leaves an FPRP (the false positive report rate) of 31%. So improving power would still leave a big problem. It is true that when we give more realistic values for power, say 50%, then the problem becomes even worse, with an FPRP of 47%. Still, most of this 47% is account for by the value of α. (For very low powered studies, the influence of low power on the FPRP does become more significant.) Increasing power of studies in fields where power is often low is certainly valuable. But it is no panacea.

4.2 Publication bias

Publication bias is any kind of bias where the details of the results of a trial affect whether it is published (independently of other factors in the research, such as quality). The principal form of publication bias is a tendency to publish research that shows a statistically significant outcome more than studies with negative results. Such bias can be exercised by the journal (which declines to publish negative results) or by the researcher (who declines to submit them—the so-called file-drawer problem (Rosenthal [1979])). This is the form of publication bias that I will assume is at work, as evidence suggests it is in biomedical research and in psychology (Ross

et al. [2012]). Publication bias has been cited as one source of the replication crisis (Pashler and Wagenmakers [2012]; Francis [2012]; Ferguson and Heene [2012]; Romero [2016]).

Publication bias on its own cannot be an explanation of the replication crisis. Imagine circumstances where there is strong publication bias at work—only positive results get published. But the scientists and journals also apply very high standards of quality and set a high bar for statistical significance. And let's say also that hypotheses are tested only when they have solid theoretical backing. Since a powerful filter for eliminating false hypotheses is being applied, it will be the case that when one sees a report of a positive result, that positive result is very likely to be true. There will few false positives, hence few failed replications. On the other hand, when one formulates a hypothesis on which there is no published information, one will not know whether that is because the hypothesis has been tested and found unproven or false (an unpublished negative result) or because the hypothesis has not been tested. One's ignorance in the latter case does not, however, undermine one's knowledge in the positive case.

So publication bias cannot explain the replication crisis because it cannot explain why there are any false positive results. And without false positives there can be no crisis. But publication bias can exacerbate the crisis. If there is a non-negligible chance of a false positive result being published, then it will be important that negative results on the same topic also get published. If negative outcomes are published, then the community is less likely to accept the positive result as proven, because, for example, a meta-analysis shows that when all the available evidence is pooled, the effect is not statistically significant. Publication bias renders meta-analysis less reliable (Colhoun *et al.* [2003]; Ferguson and Heene [2012]).[15] Or if the negative result is published first, then in order to claim a positive outcome subsequent researchers will need to carry out trials of greater size. That very fact will itself mean that the publication of a false positive is reduced. Conversely, bias preventing the publication of negative results will make it more likely that false positives will be accepted as true.

Publication bias is a serious problem for science, one which the AllTrials initiative is attempting to address. It does contribute to the replication crisis, but only marginally. The problem with the replication crisis is the publication of erroneous studies, not the non-publication of

[15]Though, as Romero ([2016], p. 65) points out, if the null hypothesis is true, then in the long run published positive results in one direction should be cancelled out by positive results in the opposite direction.

correct studies.

4.3 Bias, questionable research practices, and fraud

A large part of the sense that there is a crisis in some parts of science is the concern that the best explanation of the many replication failures is that many scientists are carrying out badly managed experiments that allow for unconscious bias, or are using self-serving analytic techniques such as *p*-hacking, or are even engaging in conscious fraud.[16] The culture of 'publish or perish' is widely blamed for poor standards and practices (John *et al.* [2012]; Simmons *et al.* [2011])—'questionable research practices', as they have become known—along with journal editors' biases towards positive over negative results (see above) and in favour of papers meeting inappropriate aesthetic standards (Giner-Sorolla [2012]). There are powerful incentives for researchers to produce interesting positive results and so it would be no surprise that some researchers cut corners, leading to outcomes that cannot later be replicated (Romero [2017]), as well as disincentivizing replication itself.

There is no doubt that there are cases of this kind. Most notoriously, the social psychologist Diederik Stapel was found by Tilburg University to have committed fraud on a huge scale in the first decade of this century, leading to the retraction of over fifty publications in social psychology. The report produced by Tilburg University argued that the scientific community had itself failed, in that it was insufficiently critical and too ready to accept results that confirmed researchers' intuitions and expectations.[17] Stapel is at one extreme. But it is easy to see how others might unconsciously allow themselves to bias their experiments (as is shown, ironically, by many results in cognitive and social psychology).

That some false positives are to be explained in this way is clear. But what proportion? To answer that reliably would require a careful examination of studies that have failed replication attempts, looking for signs of bias, poor research practices, and fraud. There is some evidence on this, but not enough yet to draw a clear conclusion. It may be that there are differences in this respect between different fields. Research teams in social psychology tend to be rather

[16] *p*-hacking occurs when researchers have collected data on many variables, look for correlations in the data for which $p < 0.05$, and then report these as statistically significant.

[17] A plausible conclusion but one that has been challenged by social psychologists.

smaller than those engaged in biomedical research and for that reason it would be easier for poor practice to go unchallenged and unchecked in the former than in the latter. It is not possible to draw a definitive conclusion in comparing the poor practice explanation of the 'crisis' with the one I have presented. However, the plausibility of the explanation presented in this paper does mean that we should be circumspect before resorting to moral panic. There is another explanation of the replication crisis that is consistent with experiments being carried out to impeccable standards. With low π and non-negligible α even excellent scientific practice will yield many false positives and hence many failed replications.

I note, however, that accepting α as high as 5% might be regarded as a questionable research practice of the community as a whole. The research community, and its most successful practitioners in particular, benefit from a relatively high α. They are incentivized to publish. A lower α would required larger sample sizes and that means more expensive experiments. Which in turn means fewer experiments and so fewer publications. And so successful scientists in these fields have no reason to pursue greater PPV.

5 What Is to Be Done?

In the light of the forgoing how should science respond to the replication crisis? There are three possibilities: (1) Do nothing—this is the nature of these sciences at their current stage of development. (2) Seek means of generating research hypotheses that are more likely to be true. (3) Require that the experimental research be of even higher quality, in particular by requiring a rather lower α than the 0.05 currently regarded as acceptable.

5.1 Quietism: This is the nature of science

The quietist approach, (1), proposes that we should just accept that it is in the nature of science that we get things wrong, and that this is particularly true with sciences in early stages of development. Popper ([1963]), for example, urged scientists to formulate bold hypotheses. But bold hypotheses are likely to be false. So we should, like Popper himself, expect to find that

our hypotheses, although they may pass some tests, will be falsified in due course.[18] Thus a corollary of the quietist position is that one should have a corresponding lowered credence even in hypotheses that have passed the tests that we have set them. That in turn should influence how think of new hypotheses. In the light of the falsity feedback effect, one should be wary of placing too much prior confidence in new hypotheses that are modelled on other, apparently successful hypotheses. The quietist must accept that the difference between passing and failing a single test does not correlate that closely with the difference between truth and falsity. Consequently the quietist should value replication studies much more than they are currently valued in many areas of science. Klein ([2014], p. 327), for example, reports that the *Journal of Personality and Social Psychology* does not publish replication studies as a mater of policy, even when the replication concerns an alleged finding of considerable significance.[19] The quietist should deplore this.

Quietism may be a reasonable stance to take in social psychology. The quietist may hope that in due course certain results will indeed stand out as reliable (thanks to successful replications) and that a stronger science may crystallize around these. And that may assist the development of a strong theory in the field. Matters may be thought to be different in clinical medicine. For here it is important that we know which hypotheses about the effectiveness of an intervention are true and which are false. If a false hypothesis passes a conventionally accepted test (such as a phase III randomized controlled trial with a 95% confidence interval), then that may initiate the use of that intervention (subject to various regulatory conditions). Not only will the intervention (for example, a new drug) fail to do good, it will do harm. Most treatments have harmful side-effects that are acceptable only because they are outweighed by the supposed benefits. Furthermore, use of one treatment will often preclude or displace the use of another treatment. So a patient who takes a new drug that although ineffective has achieved a positive result in an RCT may thereby be missing out on a treatment that really is effective (even if at a level lower than that falsely claimed for the new drug). And belief that this treatment is

[18]Popper's reasons for scepticism are, however, not those discussed in this paper, but lie instead in his rejection of induction.

[19]For example, that journal's editor declined to publish replications (with null results) of Bem's ([2011]) findings concerning precognition (Ritchie *et al.* [2012]).

effective will discourage investment in developing further new treatments.[20]

5.2 Higher quality hypotheses

The second approach advocates trying to generate hypotheses with a greater probability of truth. This is, of course, rather easier said than done. In medicine this amounts to reaffirming the importance of basic research. It may also encourage the continuation of basic research beyond the point when a new therapeutic idea becomes plausible. Rather than moving straight to a trial, as in the case of Bapineuzumab, further basic research may be a better use of resources, so that hypothesized mechanisms of action can be better confirmed before being used to generate or support a proposal regarding an intervention. In some cases greater use of observational data may be used to supplement mechanistic hypotheses. In social psychology matters are different. Here the problem is not so much an incomplete basic theory, but lack of any basic theory at all; and the prospect of one capable of generating high quality hypotheses is some way off. In any case the route to that theory will most likely be via the collection first of a set of reliable results of the general kind that social psychology is currently trying to produce. So it will be difficult directly to improve the quality of hypotheses in social psychology. However, it might be that over time a greater proportion of the accepted hypotheses are true, as replication weeds out the false positives and more stringent tests (see below) prevent them from arising in the first place. That being the case the falsity feedback effect with be increasingly dominated by a truth feedback effect, as the models for new hypotheses are increasingly true positives rather than false positives.

5.3 Decrease α

The third solution suggested by my analysis is that we should demand higher confidence levels before we accept a result in these fields. That a result is regarded as statistically significant when its p-value is less than 0.05 is a convention, one widely adopted in medicine and psychology, but

[20]The point about the pragmatic implications of quietism does have some relevance to social psychology too, since many of its results have been used in devising policy or in the business of professional development and training personnel.

not in other sciences.[21] Earlier I mentioned that physics requires outcomes that are more than 5σ from the mean before a null hypotheses is definitively rejected, whereas in biomedicine and social psychology 1.96σ is the norm. (1.96σ is equivalent to a confidence interval of 95% in a two-tailed test.) A 95% confidence interval looks good, but, as our discussion of the the base rate fallacy makes clear, a 95% confidence interval does not lead to a probability of 95% that a hypothesis is true given that is passes the test. To make that inference is to commit the *p*-value fallacy. To get a PPV of 95%, when $\pi = 0.1$, we would need to reduce α to almost 0.005 (to increase our confidence level to 2.85σ, to be precise). As opposed to increasing power alone, decreasing α does, in principle, allow us to increase PPV to any desired value (see Figure 2).

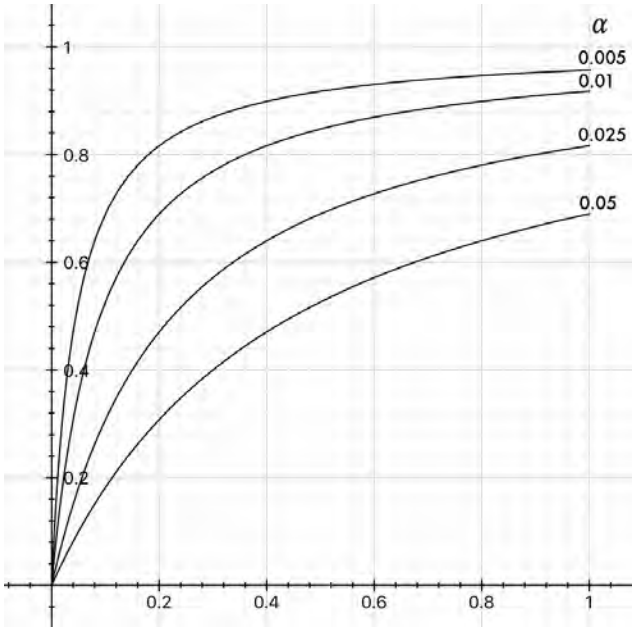

Figure 2: PPV (*y*-axis) as a function of power (*x*-axis) when $\pi = 0.1$, for different values of α.

Others have also called for lower α. Johnson ([2013]) and Benjamin *et al.* ([2018]) suggest that *p*-values should be below 0.005 before a result is accepted as statistically significant.[22] One of the motivations for Benjamin *et al.* is that this change would help address the replication crisis. Others resist such a move for various reasons. Crane ([unpublished]), for example,

[21] It is unclear why the 5% significance level is regarded as appropriate. We find it first clearly stated in Fisher, but its origins go back further (Cowles and Davis [1982]). I suspect that it is widely accepted because a one-in-twenty occurrence sounds rare enough. But, as discussed, it is a fallacy to assume that this means that one in twenty positive results are erroneous.

[22] Yet others call for the abandonment of null hypothesis significance testing altogether (for example, McShane *et al.* [unpublished]). Bayesians in particular propose alternatives (for example, Masson [2011]).

asserts 'P-hacking and the reproducibility crisis: like smoking and lung cancer, one cannot be discussed without the other' and so Benjamin *et al.* fail to address the central issue of the crisis. But that assertion depends on assuming that the central cause of the crisis is the prevalence of questionable research practices. While *p*-hacking and the like contribute to the problem, this analysis suggest that we cannot safely assume that poor practice is the principle contributor. Wacholder *et al.* criticize a similar proposal from Colhoun *et al.* ([2003]) for lowering α:

> We consider setting a low α level to be an indirect and inferior means to achieve the desired end of a low FPRP, because an FPRP can be high even for a low observed P value when the prior probability is low. Moreover, insisting on a very low P value before any finding is considered statistically significant may unnecessarily reduce statistical power when the prior probability is high, thereby constraining research on diseases with rare genetic variants or on diseases for which studies with large sample sizes are unrealistic. (Wacholder *et al.* [2004])

However, what these comments show is that there is no single α value that is appropriate for all fields. We have seen that physics works with very low α indeed. For research on carefully targeted hypotheses that are backed by strong biological rationale, then the $\alpha = 0.05$ standard may remain appropriate. The point about orphan diseases is an important one. If there are too few patients to support research with a very low α (see below), then it will be necessary to ensure that only those hypotheses are tested that are well-motivated by strong underlying theory. But if that is not possible either, then we have to accept that the results of some such research will remain tentative. Benjamin *et al.* ([2018]) propose that results with $0.005 < p < 0.05$ should be called 'suggestive'.

It is true that one consequence of this proposal is that trials need to be much larger if they are to detect the same effects while being statistically significant at lower significance levels, such as at the 0.5% level ($\alpha = 0.005$). That will likely be too demanding for psychology. It may not be quite such an obstacle in clinical medicine where quite large trials are already common. A current advantage of large trials for pharmaceutical companies is that they are able to detect very small effects that are statistically significant at the standard 5% level. Reducing α would mean that many of these results are no longer counted as statistically significant. That itself

may, paradoxically, be a benefit, since many such small effects, while statistically significant at the 5% level are not clinically significant. For example, Bhardwaj *et al.* ([2004]) discuss a large trial (2209 patients) of penciclovir, a topical treatment for herpes labialis (cold sores). There was a median reduction in time to healing of lesions from 5.5 days in the placebo group to 4.8 days in the treatment group. This was statistically significant at the 1% level ($p < 0.001$). But as the authors report, 'the results of this study, while statistically significant, lack much clinical relevance [...] This study, by using a very large sample size, detected a difference so small that it is probably not of much clinical benefit to patients'. That is, while a reduction in α will lose many false positive results it will also lose some true positive results, but these will often be results of limited or no clinical relevance. The demand for more stringent statistical standards and so for larger trials may mean that there will be fewer trials. Again, that need not be a disadvantage in the context of a large number of false positives. Fewer trials will mean that resources are directed towards testing those hypotheses with the best chances of producing a positive result (those thought most likely to be true) rather than trials that are little better than fishing expeditions. That in turn will itself improve the PPV.

6 Conclusion

Let us return to a case such as that discussed in the Harvard Medical School study. Consider subject S who assumes that a test for some disease is highly reliable—it is 95% accurate. S then unwittingly falls victim to the base rate fallacy and so expects 19 out of 20 people with a positive test result to be suffering from the disease. S then discovers that when retested the majority of those who had a positive test result now come back with a negative result. S is now likely to think she make a mistaken assumption—the test was not as reliable as she had assumed. In fact, it seems to be downright unreliable. S may lose faith in that test and perhaps begin to doubt other aspects of modern medicine. But that would not be the correct conclusion to draw, as we have seen. Things only look that way because S committed the fallacy of base rate neglect.

This paper has argued that we can explain the replication crisis and the scientific community's concern about it in the same way. By focusing on an α value of 0.05 we think we ought

to be able to have a high level of confidence that a hypothesis is true, if it is confirmed by a properly conducted study will be true. On subsequently finding grounds, such as failure to replicate, for holding that many such hypotheses are false, we are inclined to think that the original studies were not properly conducted. The belief that many studies in psychology or clinical medicine are not properly conducted takes us a long way towards a crisis of confidence in those sciences. This paper argues that this is fallacious reasoning. To say this is not to deny that questionable research practices have contributed to the existence of false positives, or that the incentive structures of science can sometimes act against the search for truth and replicable findings (Nosek *et al.* [2012]; Higginson and Munafò [2016]; Romero [2016]; Romero [2017]; Christian [2017]). These are undoubtedly real problems for science. Nonetheless, the human tendency to attribute intention, malfeasance, and defect of character where we see harm (Jones and Harris [1967]; Ross [1977]; Harman [1999]; Knobe and Burra [2006]), such as publishing unreproducible findings, may lead us to overestimate their significance. It is important to be aware that even well-performed research can be quite likely to produce a false positive result. Recognizing this might reduce the acrimony and tension that surrounds failed replication.

The analysis I have given lends itself very naturally to a Bayesian treatment (see Colhoun *et al.* [2003], p. 869; Wacholder *et al.* [2004], p. 439). I do not pursue the details here because it is clear that the conclusions are exactly the same. In passing I note the following: First, what we have called π here is the frequency with which a means of generating hypotheses generates true hypotheses. It could be interpreted as a credence, a degree of belief. If that degree of belief is determined by the track record of hypotheses in the field, as the objective Bayesian would recommend (Williamson [2010]), then the Bayesian analysis agrees with that given here. Second, it is important to note that the frequentist is entitled to consider the source of the hypothesis (and so the value of π) in deciding how the evidence bears on a hypothesis; the frequentist is not obliged to reject the null hypothesis if the p-value obtained from an experiment is less than 0.05. Fisher ([1934], p. 3) himself refers to 'the salutary habit of repeating important experiments, or of carrying out original observations in replicate'. He also emphasized the benefit of replicating experiments to increase the confidence we have in an outcome, stating that the 'confidence to be placed in a result depends not only on the magnitude of the mean

value obtained, but equally on the agreement between parallel experiments' ([1934], p. 123) .

Acknowledgements

I am grateful to Zoe Fritz, Alison Hills, Richard Holton, Marcus Munafò, Jan Sprenger, anonymous referees for the BJPS, and audiences in Bristol, Oxford, King's College London, and Edinburgh (BSPS annual conference) for advice, comments, and discussion.

Department of Philosophy

King's College London

London UK

Alexander.Bird@kcl.ac.uk

References

Baker, M. [2016]: 'Is There a Reproducibility Crisis?', *Nature*, **533**, pp. 452–4.

Bateson, M., Nettle, D. and Roberts, G. [2006]: 'Cues of Being Watched Enhance Cooperation in a Real-World Setting', *Biology Letters*, **2**, pp. 412–14.

Begley, C. G. and Ellis, L. M. [2012]: 'Drug Development: Raise Standards for Preclinical Cancer Research', *Nature*, **483**, pp. 531–3.

Bem, D. J. [2011]: 'Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect', *Journal of Personality and Social Psychology*, **100**, pp. 407–25.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A., Hadfield, J. D., Hedges, L. V., Held, L., Ho, T. H., Hoijtink, H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Jones, J. H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M.,

Moore, D., Morgan, S. L., Munafò, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J. and Johnson, V. E. [2018]: 'Redefine Statistical Significance', *Nature Human Behaviour*, **2**, pp. 6–10.

Bhardwaj, S. S., Camacho, F., Derrow, A., Fleischer, A. B. and Feldman, S. R. [2004]: 'Statistical Significance and Clinical Relevance: The Importance of Power in Clinical Trials in Dermatology', *Archives of Dermatology*, **140**, pp. 1520–3.

Blanchette, I. and Dunbar, K. [2000]: 'How Analogies Are Generated: The Roles of Structural and Superficial Similarity', *Memory and Cognition*, **29**, pp. 730–5.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J. and Munafò, M. R. [2013]: 'Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience', *Nature Reviews Neuroscience*, **14**, pp. 365–76.

Carbon, C.-C. and Hesslinger, V. M. [2011]: 'Bateson *et al*.'s (2006) Cues-of-Being-Watched Paradigm Revisited', *Swiss Journal of Psychology*, **70**, pp. 203–10.

Carney, D. R., Cuddy, A. J. C. and Yap, A. J. [2010]: 'Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance', *Psychological Science*, **21**, pp. 1363–8.

Casscells, W., Schoenberger, A. and Graboys, T. [1978]: 'Interpretation by Physicians of Clinical Laboratory Results', *New England Journal of Medicine*, **299**, pp. 999–1001.

Christian, A. [2017]: 'On the Suppression of Medical Evidence', *Journal for General Philosophy of Science*, **48**, pp. 395–418.

Cohen, J. [1990]: 'Things I Have Learned (So Far)', *American Psychologist*, **45**, pp. 1304–312.

Colhoun, H., McKeigue, P. and Davey Smith, G. [2003]: 'Problems of Reporting Genetic Associations with Complex Outcomes', *The Lancet*, **361**, pp. 865–72.

Cowles, M. and Davis, C. [1982]: 'On the Origins of the .05 Level of Statistical Significance', *American Psychologist*, **37**, pp. 553–8.

Crane, H. [unpublished]: 'Why "Redefining Statistical Significance" Will Not Improve Reproducibility and Might Make the Replication Crisis Worse', available at <arxiv.org/abs/1711.07801>.

Doyen, S., Klein, O., Pichon, C.-L. and Cleeremans, A. [2012]: 'Behavioral Priming: It's All in the Mind, But Whose Mind?', *PLOS ONE*, **7**, p. e29081.

Dumas-Mallet, E., Button, K. S., Boraud, T., Gonon, F. and Munafò, M. R. [2017]: 'Low statistical power in biomedical science: a review of three human research domains', *Royal Society Open Science*, **4**, available at <dx.doi.org/10.1098/rsos.160254>.

Dunbar, K. [1996]: 'How Scientists Really Reason', in R. Sternberg and J. Davidson (*eds*), *The Nature of Insight*, Cambridge, MA: MIT Press, pp. 365–95.

Ferguson, C. J. and Heene, M. [2012]: 'A Vast Graveyard of Undead Theories: Publication Bias and Psychological Science's Aversion to the Null', *Perspectives on Psychological Science*, **7**, pp. 555–61.

Fisher, R. A. [1934]: *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd.

Flore, P. C. and Wicherts, J. M. [2015]: 'Does Stereotype Threat Influence Performance of Girls in Stereotyped Domains? A Meta-analysis', *Journal of School Psychology*, **53**, pp. 25–44.

Francis, G. [2012]: 'Publication Bias and the Failure of Replication in Experimental Psychology', *Psychonomic Bulletin and Review*, **19**, pp. 975–91.

Funk, C. and Kennedy, B. [2016]: 'The Politics of Climate', Tech. rep., Pew Research Center.

Giner-Sorolla, R. [2012]: 'Science or Art? How Aesthetic Standards Grease the Way through the Publication Bottleneck but Undermine Science', *Perspectives on Psychological Science*, **7**, pp. 562–71.

Goodman, S. and Greenland, S. [2007]: 'Why Most Published Research Findings Are False: Problems in the Analysis', *PLOS Medicine*, **4**, p. e168.

Goodman, S. N. [1999]: 'Toward Evidence-Based Medical Statistics, 1: The *P* Value Fallacy', *Annals of Internal Medicine*, **130**, pp. 995–1004.

Hagen, R. L. [1997]: 'In Praise of the Null Hypothesis Statistical Test', *American Psychologist*, **52**, pp. 15–24.

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., Ridder, D. T. D. D., Dewitte, S., Elson, M., Evans, J. R., Fay, B. A., Fennis, B. M., Finley, A., Francis, Z., Heise, E., Hoemann, H., Inzlicht, M., Koole, S. L., Koppel, L., Kroese, F., Lange, F., Lau, K., Lynch, B. P., Martijn, C., Merckelbach, H., Mills, N. V., Michirev, A., Miyake, A., Mosser, A. E., Muise, M., Muller, D., Muzi, M., Nalis, D., Nurwanti, R., Otgaar, H., Philipp, M. C., Primoceri, P., Rentzsch, K., Ringos, L., Schlinkert, C., Schmeichel, B. J., Schoch, S. F., Schrama, M., Schütz, A., Stamos, A., Tinghög, G., Ullrich, J., vanDellen, M., Wimbarti, S., Wolff, W., Yusainy, C., Zerhouni, O. and Zwienenberg, M. [2016]: 'A Multilab Preregistered Replication of the Ego-Depletion Effect', *Perspectives on Psychological Science*, **4**, pp. 546–73.

Harman, G. [1999]: 'Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error', *Proceedings of the Aristotelian Society*, **99**, pp. 315–31.

Higginson, A. D. and Munafò, M. R. [2016]: 'Current Incentives for Scientists Lead to Underpowered Studies with Erroneous Conclusions', *PLOS Biology*, **14**.

Ioannidis, J. P. A. [2005a]: 'Contradicted and Initially Stronger Effects in Highly Cited Clinical Research', *Journal of the American Medical Association*, **294**, pp. 218–28.

Ioannidis, J. P. A. [2005b]: 'Why Most Published Research Findings Are False', *PLOS Medicine*, **2**.

John, L. K., Loewenstein, G. and Prelec, D. [2012]: 'Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling', *Psychological Science*, **23**, pp. 524–32.

Johnson, V. E. [2013]: 'Revised Standards for Statistical Evidence', *PNAS*, **110**, pp. 19313–17.

Johnson, V. E., Payne, R. D., Wang, T., Asher, A. and Mandal, S. [forthcoming]: 'On the Reproducibility of Psychological Science', *Journal of the American Statistical Association*, **112**, pp. 1–10.

Jones, E. E. and Harris, V. A. [1967]: 'The Attribution of Attitudes', *Journal of Experimental Social Psychology*, **3**, pp. 1–24.

Klein, R. A., Ratliff, K. A., Vianello, M., Jr., R. B. A., Štěpán Bahník, Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., Hasselman, F., Hicks, J. A., Hovermale, J. F., Hunt, S. J., Huntsinger, J. R., IJzerman, H., John, M.-S., Joy-Gaba, J. A., Kappes, H. B., Krueger, L. E., Kurtz, J., Levitan, C. A., Mallett, R. K., Morris, W. L., Nelson, A. J., Nier, J. A., Packard, G., Pilati, R., Rutchick, A. M., Schmidt, K., Skorinko, J. L., Smith, R., Steiner, T. G., Storbeck, J., Swol, L. M. V., Thompson, D., van't Veer, A. E., Vaughn, L. A., Vranka, M., Wichman, A. L., Woodzicka, J. A. and Nosek, B. A. [2014]: 'Investigating Variation in Replicability', *Social Psychology*, **45**, pp. 142–52.

Klein, S. B. [2014]: 'What Can Recent Replication Failures Tell Us about the Theoretical Commitments of Psychology?', *Theory and Psychology*, **24**, pp. 326–38.

Knobe, J. and Burra, A. [2006]: 'The Folk Concepts of Intention and Intentional Action: A Cross-cultural Study', *Journal of Cognition and Culture*, **6**, pp. 113–32.

Kuhn, T. S. [1970]: *The Structure of Scientific Revolutions*, Chicago, IL: University of Chicago Press.

Lakens, D. [2014]: 'The Null Is Always False (Except When It Is True)', available at <daniellakens.blogspot.co.uk/2014/06/the-null-is-always-false-except-when-it.html>.

Leung, A. K., Kim, S., Polman, E., Ong, L. S., Qiu, L., Goncalo, J. A. and Sanchez-Burks, J. [2012]: 'Embodied Metaphors and Creative "Acts"', *Psychological Science*, **23**, pp. 502–9.

Masson, M. E. J. [2011]: 'A Tutorial on a Practical Bayesian Alternative to Null-Hypothesis Significance Testing', *Behavior Research Methods*, **43**, pp. 679–90.

McShane, B. B., Gal, D., Gelman, A., Robert, C. and Tackett, J. L. [unpublished]: 'Abandon Statistical Significance', available at <arxiv.org/abs/1709.07588>.

Meehl, P. E. [1978]: 'Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology', *Journal of Consulting and Clinical Psychology*, **46**, pp. 806–34.

Meehl, P. E. [1990]: 'Why Summaries of Research on Psychological Theories are Often Uninterpretable', *Psychological Reports*, **66**, pp. 195–244.

Nosek, B. A., Spies, J. R. and Motyl, M. [2012]: 'Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth over Publishability', *Perspectives on Psychological Science*, **7**, pp. 615–31.

Oostenbroek, J., Suddendorf, T., Nielsen, M., Redshaw, J., Kennedy-Costantini, S., Davis, J., Clark, S. and Slaughter, V. [2016]: 'Comprehensive Longitudinal Study Challenges the Existence of Neonatal Imitation in Humans', *Current Biology*, **26**, pp. 1334–8.

Pashler, H. and Wagenmakers, E.-J. [2012]: 'Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence?', *Perspectives on Psychological Science*, **7**, pp. 528–30.

Popper, K. [1959]: *The Logic of Scientific Discovery*, London: Hutchinson.

Popper, K. [1963]: *Conjectures and Refutations: The Growth of Scientific Knowledge*, London: Routledge and Kegan Paul.

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S. and Weber, R. A. [2015]: 'Assessing the Robustness of Power Posing', *Psychological Science*, **26**, pp. 653–6.

Ritchie, S. J., Wiseman, R. and French, C. C. [2012]: 'Replication, Replication, Replication', *The Psychologist*, **25**, pp. 346–8.

Romero, F. [2016]: 'Can the Behavioral Sciences Self-correct? A Social Epistemic Study', *Studies in History and Philosophy of Science*, **60**, pp. 55–69.

Romero, F. [2017]: 'Novelty vs. Replicability: Virtues and Vices in the Reward System of Science', *Philosophy of Science*, **84**, pp. 1031–43.

Rosenthal, R. [1979]: 'The File Drawer Problem and Tolerance for Null Results', *Psychological Bulletin*, **86**, pp. 638–41.

Ross, J. S., Tse, T., Zarin, D. A., Xu, H., Zhou, L. and Krumholz, H. M. [2012]: 'Publication of NIH Funded Trials Registered in ClinicalTrials.gov: Cross-sectional Analysis', *BMJ*, **344**, available at <www.bmj.com/content/344/bmj.d7292>.

Ross, L. [1977]: 'The Intuitive Psychologist and His Shortcomings', in L. Berkowitz (*ed.*), *Advances in Experimental Social Psychology*, New York: Academic Press.

Sanna, L. J., Chang, E. C., Miceli, P. M. and Ludberg, K. B. [2011]: 'Rising up to Higher Virtues: Experiencing Elevated Physical Height Uplifts Prosocial Actions', *Journal of Experimental Social Psychology*, **47**, pp. 417–6.

Simmons, J. P., Nelson, L. D. and Simonsohn, U. [2011]: 'False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant', *Psychological Science*, **22**, pp. 1359–66.

Stepper, S. and Strack, F. [1993]: 'Propioceptive Determinants of Emotional and Nonemotional Feelings', *Journal of Personality and Social Psychology*, **64**, pp. 211–20.

Sterne, J. and Davey Smith, G. [2001]: 'Sifting the Evidence—What's Wrong with Significance Tests?', *Physical Therapy*, **81**, pp. 1464–9.

Strack, F., Martin, L. L. and Stepper, S. [1988]: 'Inhibiting and Facilitating Conditions of the Human Smile: A Nonobtrusive Test of the Facial Feedback Hypothesis', *Journal of Personality and Social Psychology*, **54**, pp. 768–77.

Szucs, D. and Ioannidis, J. P. A. [2017]: 'Empirical Assessment of Published Effect Sizes and Power in the Recent Cognitive Neuroscience and Psychology Literature', *PLOS Biology*, **15**.

Wacholder, S., Chanock, S., Garcia-Closas, M., El ghormli, L. and Rothman, N. [2004]: 'Assessing the Probability That a Positive Report Is False: An Approach for Molecular Epidemiology Studies', *Journal of the National Cancer Institute*, **96**, pp. 434–42.

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., and Gronau, Q. F. [2016]: 'Registered Replication Report: Strack, Martin, and Stepper (1988)', *Perspectives on Psychological Science*, **11**, pp. 917–28.

Williamson, J. [2010]: *In Defence of Objective Bayesianism*, Oxford: Oxford University Press.