**On the evidentiary emptiness of failed replications**

Jason Mitchell
Harvard University
1 July 2014

- Recent hand-wringing over failed replications in social psychology is largely pointless, because unsuccessful experiments have no meaningful scientific value.
- Because experiments can be undermined by a vast number of practical mistakes, the likeliest explanation for any failed replication will always be that the replicator bungled something along the way. Unless direct replications are conducted by flawless experimenters, nothing interesting can be learned from them.
- Three standard rejoinders to this critique are considered and rejected. Despite claims to the contrary, failed replications do not provide meaningful information if they closely follow original methodology; they do not necessarily identify effects that may be too small or flimsy to be worth studying; and they cannot contribute to a cumulative understanding of scientific phenomena.
- Replication efforts appear to reflect strong prior expectations that published findings are not reliable, and as such, do not constitute scientific output.
- The field of social psychology can be improved, but not by the publication of negative findings. Experimenters should be encouraged to restrict their "degrees of freedom," for example, by specifying designs in advance.
- Whether they mean to or not, authors and editors of failed replications are publicly impugning the scientific integrity of their colleagues. Targets of failed replications are justifiably upset, particularly given the inadequate basis for replicators' extraordinary claims.

## *The sociology of scientific failure*

When we expect an experiment to yield certain results, and yet it fails to do so, scientists typically work to locate the source of the failure. In principle, the cause of an experimental failure could lurk anywhere; philosophers have pointed out that a failed experiment might very well indicate a previously undetected flaw in our system of logic and mathematics[1]. In practice, however, most scientists work from a mental checklist of likely culprits. At the top of this list, typically, are "nuts-and-bolts" details about the way in which the experiment was carried out—was my apparatus working properly?; did my task operationalize the variable I was aiming for?; did I carry out the appropriate statistical analysis in the correct way?; and so on. Very often, the source of the failure is located here, if only because the list of practical mistakes that can undermine an experiment is so vast.

Considerably lower down the list are various doubts for expecting particular results in the first place, such as uncertainty about the theory that predicted them or skepticism about reports of similar effects[2]. In other words, when an experiment fails, scientists typically first assume that they bungled the details of the experiment before concluding that something must be wrong with their initial reasons for having conducted it in the first place (or that logic and mathematics suffer some fatal flaw). This makes good sense: it would be inane to discard an entire theoretical edifice because of one researcher's undetected copy-and-paste error or other such practical oversight. In my own research, I have made many mistakes that initially went unnoticed. I have, for instance, belatedly realized that a participant was earlier run in a similar pilot version of the experiment and already knew the hypotheses; I've inadvertently run analyses on a dozen copies of the same set of fMRI images instead of using different data for each subject; I have written analysis code that incorrectly calculated the time of stimulus onset; and on and on. I might be embarrassed by a full accounting of my errors, except for the fact that I'm in good company—every other scientists I know has experienced the same frequent failings, which is why the first, second, and third best explanation of any failed effect has always been that I mucked something up along the way.

Consider how recent replication efforts invert these assumptions, however. A replication attempt starts with good reasons to run an experiment: some theory predicts positive findings, and such findings have been reported in the literature, often more than once. Nevertheless, the experiment fails. In the normal course of science, the presumption would be that the researcher flubbed something important (perhaps something quite subtle) in carrying out the experiment, because that is far-and-away the most likely cause of a scientific failure. But if an experiment fizzles merely because of practical defects from which we all suffer, then there is nothing to be learned from a failed replication. Yes, it could be that the original effect was the result of *p*-hacking or fraud or some other scientific ugliness. Or, then again, maybe the failed experimenters just didn't quite execute perfectly.

To put a fine point on this: if a replication effort were to be capable of identifying empirically questionable results, it would have to employ flawless experimenters. Otherwise, how do we identify replications that fail simply because of undetected experimenter error? When an experiment succeeds, we can celebrate that the phenomenon survived these all-too-frequent shortcomings. But when an experiment fails, we can only wallow in uncertainty about whether a phenomenon simply does not exist or, rather, whether we were just a bit too human that time around. And here is the rub: if the most likely explanation for a failed experiment is simply a mundane slip-up, and the replicators are themselves not immune to making such mistakes, then the replication efforts have no meaningful evidentiary value outside of the very local (and uninteresting) fact that Professor So-and-So's lab was incapable of producing an effect.

This should be immediately apparent by the co-existence of both successful and failed replications. The recent special issue of *Social Psychology*, for example, features one paper that successfully reproduced observations that Asian women perform better on mathematics tests when their Asian identity, rather than their female identity, is primed.

A second paper, following the same methodology, failed to find this effect (Moon & Roeder, 2014); in fact, the 95% confidence interval does not include the original effect size. These oscillations should give serious pause to fans of replicana. Evidently, not all replicators can generate an effect, *even when that effect is known to be reliable*. On what basis should we assume that other failed replications do not suffer the same unspecified problems that beguiled Moon and Reoder? The replication effort plainly suffers from a problem of false negatives.

*The problem with recipe-following (response to rejoinder #1)*

There are three standard rejoinders to these points. The first is to argue that because the replicator is closely copying the method set out in an earlier experiment, the original description must in some way be insufficient or otherwise defective. After all, the argument goes, if someone cannot reproduce your results when following your recipe, something must be wrong with either the original method or in the findings it generated.

This is a barren defense. I have a particular cookbook that I love, and even though I follow the recipes as closely as I can, the food somehow never quite looks as good as it does in the photos. Does this mean that the recipes are deficient, perhaps even that the authors have misrepresented the quality of their food? Or could it be that there is more to great cooking than just following what's printed in a recipe? I do wish the authors would specify how many millimeters constitutes a "thinly" sliced onion, or the maximum torque allowed when "fluffing" rice, or even just the acceptable range in degrees Fahrenheit for "medium" heat. They don't, because they assume that I share tacit knowledge of certain culinary conventions and techniques; they also do not tell me that the onion needs to be peeled and that the chicken should be plucked free of feathers before browning. If I do not possess this tacit know-how—perhaps because I am globally incompetent, or am relatively new to cooking, or even just new to cooking Middle Eastern food specifically—then naturally, my outcomes will differ from theirs.

Likewise, there is more to being a successful experimenter than merely following what's printed in a method section. Experimenters develop a sense, honed over many years, of how to use a method successfully. Much of this knowledge is implicit. Collecting meaningful neuroimaging data, for example, requires that participants remain near-motionless during scanning, and thus in my lab, we go through great lengths to encourage participants to keep still. We whine about how we will have spent a lot of money for nothing if they move, we plead with them not to sneeze or cough or wiggle their foot while in the scanner, and we deliver frequent pep talks and reminders throughout the session. These experimental events, and countless more like them, go unreported in our method section for the simple fact that they are part of the shared, tacit know-how of competent researchers in my field; we also fail to report that the experimenters wore clothes and refrained from smoking throughout the session. Someone without full possession of such know-how—perhaps because he is globally incompetent, or new to science, or even just new to neuroimaging specifically—could well be expected to bungle one or more of these important, yet unstated, experimental details. And because there are many more ways to do an experiment badly than to do one well, recipe-following will

commonly result in failure to replicate[3].

*Why it can be interesting to study flimsy effects (response to rejoinder #2)*

A second common rejoinder is to argue that if other professional scientists cannot reproduce an effect, then it is unlikely to be "real." Science should focus on robust effects that can be produced easily; any phenomenon that is difficult to observe or that "comes-and-goes" unpredictably can hardly be worth studying.

This is a slightly more seductive argument, but it, too, falls short. Many of the most robust and central phenomena in psychology started life as flimsy and capricious effects, their importance only emerging after researcher developed more powerful methods with which to study them. Perhaps the most pertinent example comes from research on implicit prejudice, the observation that perceivers unconsciously evaluate social groups as positive or negative and automatically associate members of such groups with specific stereotypes. These effects are some of the best-documented phenomena in modern social psychology, thanks largely to the development of methods, such as the Implicit Association Test, for reproducing them consistently. We might disagree about the most appropriate interpretation of these effects, but few of us doubt that they are shockingly reliable.

One might be excused, then, for assuming that it has always been so easy. But until the late 1990s, the field of implicit prejudice was pretty darn messy. Effects were difficult to obtain; they were associated with wildly fluctuating effect sizes; and researchers could not agree on whether there were important, yet-to-be-properly-understood moderators of such effects (e.g., Gilbert & Hixon, 1991). Some foundational papers reported sharp decreases in effect size over successive studies (Banaji, Rothman, & Hardin, 1993; Banaji & Hardin, 1996; Fazio, Jackson, Dunton, & Williams, 1995), and some even failed to replicate their own findings (e.g., Fazio et al., 1995). Many studies were sorely underpowered. And important aspects of these effects—such as whether individuals associated negativity with outgroup members, or just positivity with the ingroup—were reported by some labs (Dovidio, Evans, & Tyler, 1986) but not others (Gaertner & McLaughlin, 1983; Fazio et al., 1995).

If these studies had been subject to the kind of replication effort that we currently see, many would not easily replicate; they reported flimsy effects that were difficult to obtain even by seasoned veterans. But with twenty years' worth of hindsight, we know that these studies were, in fact, telling us about a highly reliable phenomenon—we just didn't have the right methods for producing it consistently. Luckily for the study of implicit prejudice, those methods were eventually located; other fields may not (yet) have access to similarly powerful tools. However, the history of implicit prejudice research makes this much clear: the fact that a scientific phenomenon is small or mercurial or difficult to obtain does not provide sufficient evidence against its genuineness or importance. Other such cases are not hard to identify[4].

*The asymmetry between positive and negative evidence (response to rejoinder #3)*

A third rejoinder argues that the replication effort ought to be considered a counterweight to our publication bias in favor of positive results. Ordinarily, negative findings are nearly impossible to publish, whereas positive findings stand a good chance of making their way into the literature. As a result, scientists usually only get to see evidence in favor of a phenomenon, even if many studies fail to observe the same effect. The replication effort addresses this disparity by allowing negative evidence to see the light of day. After all, the argument goes, if an effect has been reported twice, but hundreds of other studies have failed to obtain it, isn't it important to publicize that fact?

No, it isn't. Although the notion that negative findings deserve equal treatment may hold intuitive appeal, the very foundation of science rests on a profound asymmetry between positive and negative claims. Suppose I assert the existence of some phenomenon, and you deny it; for example, I claim that some non-white swans exist, and you claim that none do (i.e., that no swans exist that are any color other than white). Whatever our *a priori* beliefs about the phenomenon, from an inductive standpoint, your negative claim (of nonexistence) is infinitely more tenuous than mine. A single positive example is sufficient to falsify the assertion that something does not exist; one colorful swan is all it takes to rule out the impossibility that swans come in more than one color. In contrast, negative examples can never establish the nonexistence of a phenomenon, because the next instance might always turn up a counterexample. Prior to the turn of the 17$^{th}$ century, Europeans did indeed assume that all swans were white. When Dutch explorers observed black swans in Australia, this negative belief was instantly and permanently confuted. There is a striking asymmetry here: a single positive finding (of a non-white swan) had more evidentiary value than millennia of negative observations. What more, it is clear that the null claim cannot be reinstated by additional negative observations: rounding up trumpet after trumpet of white swans does not rescue the claim that no non-white swans exists. This is because positive evidence has, in a literal sense, infinitely more evidentiary value than negative evidence[5].

Thus, negative findings—such as failed replications—cannot bear against positive evidence for a phenomenon. In the same way that a parade of white swans has no evidentiary value regarding the existence or nonexistence of black swans, journals full of negative findings have no evidentiary value regarding the existence or nonexistence of scientific phenomena. Positive scientific assertion cannot be reversed solely on the basis of null observations.

Does all this mean that we have no defense against spurious claims, and must believe every reported observation, no matter how far-fetched and improbable? Not at all. Although the logic of science dictates that negative evidence can never triumph over positive evidence, we can always bring additional *positive* evidence to bear on a question. Suppose someone claims to have seen a paisley swan and insists that we must therefore abandon the belief that all swans are a single color. If I am to dislodge this claim, it won't do simply to scare up several white swans. Instead, I must provide a positive explanation for the observation: how did the report of a paisley swan come to be? For

example, I might assert that the observer is lying, or is herself deceived. I might identify faults in her method and explain how they lead to spurious conclusions. I might describe the factors that can make a swan appear as multi-colored despite being purely white. In each case, the onus is on me to make a productive assertion that accounts for the observation[6].

Only such positive evidence can bear on the reliability of social psychological effects. If one doubts a reported effect, then he must provide positive evidence for how the observation came to be. For example, he might identify some aspect of the original method that produces a spurious result. Or he might provide evidence that the original author committed some error. Or best yet, he might demonstrate that the phenomena depends on some previously unknown moderator variable. In the absence of such positive claims, however, null results have vanishingly little evidentiary value and are therefore almost never worth publicizing.

### *Why the replication efforts are not science*

The purveyors of negative results consistently present themselves as disinterested parties with no particular axe to grind, merely concerned scientific citizens whose main aim is to set the empirical record straight. The appearance of neutrality is vital to the replication project—after all, if researchers enter into the process thinking "this effect is hogwash and I'm going to show it," how do the rest of us protect against the replicators' bias (inadvertent or otherwise) towards finding negative results? Put another way, if we are to believe that positive results can be willed (or tortured) into being by those sufficiently motivated, surely we must worry that negative results can also be generated by those powerfully motivated by disbelief?

There are good reasons to conclude that the replicators are not, in actual fact, neutral with regard to the effects that they revisit, but are instead motivated by strong prior disbelief in the original findings. As we saw above, failed experiments typically trigger an attempt to locate the source of the failure; this is because researchers only conduct a particular experiment when they have strong reasons for (prior expectations about) doing so. But consider how the replication project inverts this procedure—instead of trying to locate the sources of experimental failure, the replicators and other skeptics are busy trying to locate the sources of experimental *success*. It is hard to imagine how this makes any sense unless one has a strong prior expectation that the effect does not, in fact, obtain. When an experiment fails, one will work hard to figure out why if she has strong expectations that it should succeed. When an experiment succeeds, one will work hard to figure out why to the extent that she has strong expectations that it should fail. In other words, scientists try to explain their failures when they have prior expectations of observing a phenomenon, and try to explain away their successes when they have prior expectations of that phenomenon's nonoccurrence.

Let's put this another way. Someone who publishes a replication is, in effect, saying something like, "You found an effect. I did not. One of us is the inferior scientist." I can imagine three possible conclusions to this thought. The first is that the replicator wants to

acknowledge publicly that he isn't up to snuff, that his attempt was brought low by routine mistakes; this seems an implausibly self-defeating motive, but we cannot exclude it from possibility. The second is that the replicator has no recommendation about which result to believe and simply invites us to join him in agnostic uncertainty. If so, it's hard to understand why he went through all that the effort to produce something with such little evidentiary value. The remaining conclusion is that the replicator believes—and wants us to believe—that the original finding falls short. But why would we accept that conclusion if the most likely explanation for a failed experiment is everyday human error? Unless the replicator believes that the most likely explanation for the failure is not the typical humdrum one, but something quite extraordinary, such as the original authors' incompetence or malice. A failed experiment is meaningful only if one's prior expectations favor a perfectly executed experiment over the reliability of the phenomenon. Otherwise, it merely suggests the limits of one's own experimental chops. After all, I would only tell my friends and family about the cruddy cookbook I received if I was certain that it was the authors' recipes, and not my cooking, that was lacking. There is little need to remind them of the readily-observed fact that I am capable of making mistakes.

At any rate, none of this constitutes scientific output. Science makes no progress by celebrating experimental inadequacy, nor by impishly trying to foster uncertainty. And it certainly makes no progress borne on the backs of those who make extraordinary claims (about earlier findings) on the basis of remarkably feeble evidence. If social psychologists feel the same need to justify their successes as they do their failures, then we are not doing science, but instead fighting a war of prior expectations. On one side are those who believe more strongly in their own infallibility than in the existence of reported effects; on the other are those who continue to believe that positive findings have infinitely greater evidentiary value than negative ones and that one cannot prove the null. I was mainly educated in Catholic schools, and a frequent response to impertinent children who questioned Church mysteries was simply that "for those who believe, no proof is necessary; for those who do not believe no proof is possible."[7] How strange and disconcerting that this statement provides an apt description for the current goings-on in my scientific discipline.

### *Recommendations for moving forward*

None of this is meant to imply that I think all is well in social psychology. Far from it: I think our scientific standards have been woefully low, and that we have been fooling ourselves for some time about what constitutes genuine and important effects. Our field is in desperate need of reform—it's just that replication efforts are not part of what is needed, and are likely doing more harm than good.

How should we move forward? First, it's important to distinguish between two problems plaguing the field—deliberate fraud and motivated sloppiness. We can all agree that the former is an abomination, and that anyone who fabricates data should be deposited as far away from science as possible. But tighter standards will not stop someone who ignores even the basic tenets of scientific conduct. Some amount of fraud will always be a part of

science, perpetrated by individuals who measure success in number of papers published and awards won, rather than in true discoveries made. It can, and should, be identified when possible. But fraud cannot be prevented by more exacting standards for publication, since forged results can be fabricated to any arbitrary standard. Fortunately, intentional fraud is probably very rare, in our field as in the rest of science.

On the other hand, a good deal can be done to clean up the sloppiness of our current practices. All scientists are motivated to find positive results, and social psychologists are no exception. But recent work has exposed just how easily confirmation biases of this kind can lead us to false conclusions. Many of us have allowed ourselves to be flexible about the covariates we use or the number of participants we include. We might have dimly appreciated that this flexibility was not "best practice," but papers like Simmons *et al.* (2011) illustrate just how devastating such practices can be to the search for truth. If researchers allow themselves sufficient flexibility in the collection and analysis of their data, they can produce any result, even those that are patently false. The logical conclusion is obvious: any field that does not suppress such flexibility can expect to be peppered with spurious effects.

So we must act to restrain ourselves. Merely knowing about the dangers of experimenter flexibility will itself improve our behavior; by and large, scientists are motivated by discovery of hidden truths, and want to master the methods that will reveal them. At an institutional level, authors should be encouraged to specify experimental designs in advance, and I applaud much of the effort to create a system of study preregistration (except for plans to publish null results[8]). These changes will not fix the field overnight. They will not root out whatever false positives currently inhabit the published literature. But once we require studies to be specified in advance, spurious effects will quickly disappear from the frontlines of our science. The field will have righted its course, not by reviewing its mistakes, but by instituting positive reforms for strengthening our methods of inquiry into the future.

Will these changes slow progress in social psychology? Absolutely not. That is, yes, they will probably slow the rate at which we publish, but if we cannot otherwise distinguish truth from fiction in our field, what intellectual progress are we currently making? Surely most of us would rather pick our way slowly towards the truth than run off at full tilt in some other direction.

### *The dangers of scientific innuendo*

Ted Williams once said of baseball that it's "the only field of endeavor where a man can succeed three times out of ten and be considered a good performer." He apparently never submitted an NIH grant or applied for an academic job. Science is a tough place to make a living. Our experiments fail much of the time; even the best scientists meet with a steady drum of rejections from journals, grant panels, and search committees; and most of us continually fall short of the expectations of our students, colleagues, or advisors. On the occasions that our work does succeed, we expect others to criticize it mercilessly, in public and often in our presence. And although a few scientists will land TV

appearances and become best-selling authors, most of us will find our primary reward in the work itself, in the satisfaction of adding our incremental bit to the sum of human knowledge and hoping that our ideas might manage, even if just, to influence future scholars of the mind. It takes courage and grit and enormous fortitude to volunteer for a life of this kind.

So we should take note when the targets of replication efforts complain about how they are being treated. These are people who have thrived in a profession that alternates between quiet rejection and blistering criticism, and who have held up admirably under the weight of earlier scientific challenges. They are not crybabies. What they are is justifiably upset at having their integrity questioned. Academia tolerates a lot of bad behavior—absent-minded wackiness and self-serving grandiosity top the list—but misrepresenting one's data is the unforgivable cardinal sin of science. Anyone engaged in such misconduct has stepped outside the community of scientists and surrendered his claim on the truth. He is, as such, a heretic, and the field must move quickly to excommunicate him from the fold. Few of us would remain silent in the face of such charges.

Because it cuts at the very core of our professional identities, questioning a colleague's scientific intentions is therefore an extraordinary claim. That such accusations might not be expressed directly but only whispered and hinted at hardly matters; as social psychologists, we should know better that innuendo and intimation can be every bit as powerful as direct accusation. Like all extraordinary claims, insinuations about others' scientific integrity should require extraordinary evidence. Failures to replicate do not even remotely make this grade, since they most often result from mere ordinary human failing. Replicators not only appear blind to these basic aspects of scientific practice, but unworried about how their claims affect the targets of their efforts. One senses either a profound naiveté or a chilling mean-spiritedness at work, neither of which can provide a lasting basis for reform in social psychology.

---

[1] See Quine, W.v.O. (1953) 'Two dogmas of empiricism', in *From a Logical Point of View*.

Duhem, P. (1906, tr. 1962) *The Aim and Structure of Physical Theory*, New York: Athenum. Again, Ladyman (Chapter 6).

[2] Thomas Kuhn, The Structure of Scientific Revolutions. Kuhn argues that in the course of normal science, researchers typically conduct only those experiments for which they

have strong prior expectations that certain phenomena should be observed; hence, they are generally reluctant to move down the list of plausible culprits when trying to locate a source of experimental failure.

[3] Here are a handful of the many hundreds of factors that we have learned are important to the success of our experiments but which are never explicitly described in our journal articles:

- ensure that the very edges of the screen can be seen fully by participants (many will otherwise complete the task without stopping to tell us that they could not see the stimuli)
- notice when a participant is anxious about being in the scanner, and act to decrease such anxiety
- double-check that participants can understand verbal instructions delivered through the intercom (many will otherwise just proceed without understanding the nature of their task)
- notice when a participant does not seem to fully understand the standard instructions, and re-phrase or repeat accordingly
- ensure that participants are not too cold in the scanner room, and that they do not have to go to the bathroom before the experiment begins
- ensure that room is darkened to maximize screen visibility
- secure the response box to the participant's body in a comfortable position, so that he or she does not need to adjust it during the experiment
- ensure that participant's head is not so tightly restrained to cause pain
- and so on…

[4] Another example: in the first few years of fMRI, researchers would only sometimes observe hippocampal activation during memory tasks, and when we did, the effects would often be quite underwhelming. Because neuropsychological research had already established a critical role for the hippocampus in memory formation (e.g., patient HM), we had strong reason to believe that something was wrong with our methods. Although it took some years to locate the problem, we eventually figured out that the hippocampus was engaged only by certain memory tasks, especially those that asked participants to associate two arbitrary bits of experience. Again, a genuine phenomenon spent a surprisingly long time masquerading as a flimsy empirical effect. The examples from implicit prejudice and the cognitive neuroscience of memory are just two with which I happen to be familiar; there are doubtless many more such cautionary tales.

[5] For a full treatment of these issues, see Karl Popper's *The Logic of Scientific Discovery* and *Conjectures and Refutations*. For an excellent, and thoroughly readable, review of the asymmetry between positive and negative evidence, see James Ladyman's *Understanding Philosophy of Science* (especially Chapters 2 & 3).

[6] In effect, one must turn the (potentially spurious) observation itself into a null hypothesis that can be falsified—something like "the observation of a paisley swan cannot be explained in any way other than that not all swans are monochromatic." One

can then look to refute this null hypothesis by providing evidence for one or more such alternatives.

[7] Ironically, a phrase attributed to the economist Stuart Chase.

[8] As a rule, studies that produce null results—including preregistered studies—should not be published. As argued throughout this piece, null findings cannot distinguish between whether an effect does not exist or an experiment was poorly executed, and therefore have no meaningful evidentiary value even when specified in advance. Replicators might consider a publicly-searchable repository of unpublished negative findings, but these should in no way be considered dispositive with regard to the effects of interest.