

Self-correction in science: The diagnostic and integrative motives for replication

Social Studies of Science

2021, Vol. 51 (4) 583–605

© The Author(s) 2021

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/03063127211005551

journals.sagepub.com/home/sss**David Peterson¹**  **and Aaron Panofsky¹**

Abstract

A series of failed replications and frauds have raised questions regarding self-correction in science. Metascientific activists have advocated policies that incentivize replications and make them more diagnostically potent. We argue that current debates, as well as research in science and technology studies, have paid little heed to a key dimension of replication practice. Although it sometimes serves a diagnostic function, replication is commonly motivated by a practical desire to extend research interests. The resulting replication, which we label ‘integrative’, is characterized by a pragmatic flexibility toward protocols. The goal is to appropriate what is useful, not test for truth. Within many experimental cultures, however, integrative replications can produce replications of ambiguous diagnostic power. Based on interviews with 60 members of the Board of Reviewing Editors for the journal *Science*, we show how the interplay between the diagnostic and integrative motives for replication differs between fields and produces different cultures of replication. We offer six theses that aim to put science and technology studies and science activism into dialog to show why effective reforms will need to confront issues of disciplinary difference.

Keywords

replication, reproducibility crisis, metascience, science policy

Popper (1934/2005: 45) argued that science without replication is nothing more than the fruitless cataloging of ‘occult effects’. Many commentators suggest that this is, in fact, the situation the sciences now face. Popular news outlets have picked up articles showing that only a third of experimental psychology experiments and less than 20% of cancer biology research can be repeated (Begley, 2015; Open Science Collaboration, 2015).

¹University of California Los Angeles, Los Angeles, CA, USA

Corresponding author:

David Peterson, Institute for Society and Genetics, University of California Los Angeles, Los Angeles, CA 90095, USA.

Email: davidapeterson@g.ucla.edu

A highly cited survey from *Nature* (Baker, 2016) indicates that these problems are common and widespread. First diagnosed in the social and biomedical sciences, the ‘replication crisis’ has become the motive for a metascientific social movement that is catholic in scope. Rather than problems with any individual science, they have diagnosed problems with the organizations, institutions, rules, and incentives that structure scientific exchange. The disease is systemic, and so must be the cure.

Metascientific activists have been successful at drawing both attention and resources to the issue of reproducibility. In the past five years, organizations across the US and Western Europe created to improve science and deliver us from the crisis have received over \$100 million dollars in public and private funding. This is over and above the flurry of activity initiated within universities and journals. This activism has had real consequences. Rule changes and new guidelines are being developed across scientific institutions to encourage or mandate transparency (Freese and Peterson, 2018). Mass replications are being conducted to measure the reproducibility rate of entire fields. One of these efforts was hailed as one of *Science’s* ‘breakthroughs of the year’ (Science News, 2015). More modest replication projects are being encouraged by new policies at journals.

Key to metascientists’ critiques are the arguments that (a) replications are uncommon in many fields¹ and (b) their rarity obstructs self-correction. For instance, in his article ‘Why science is not necessarily self-correcting’, biostatistician and activist Ioannidis (2012) blames systemic disincentives to replicate for allowing incorrect findings to proliferate. And, the mass replication of psychology studies conducted by the Open Science Collaboration (2015) concludes that replication provides the evidence that ‘is the scientific community’s method of self-correction and is the best available option for achieving that ultimate goal: truth’.

At odds with its universalistic rhetoric, however, the current activism around replication has emerged from a narrow category of scientist. Although a related methodological reform movement has been active in clinical research, current discussions have largely been driven by behavioral scientists, especially psychologists. Psychologists have previously played an important role in introducing new methodological standards into the sciences (Porter, 1995), and it appears they are taking the lead again. For instance, the Center for Open Science, a major recipient of funding aimed at solving the reproducibility crisis science-wide, was started by two psychologists. Because of this, it raises important questions. Will the reforms being pursued benefit all sciences? If not, does this activism represent a political movement as much as an epistemic one?

Although largely ignored within the metascientific literature, scholarship in science and technology studies (STS) has greatly complicated the picture presented by metascientists. Scholars have highlighted the importance of the ‘methodologist’ as a category of scientific actor who benefits from opening black boxes (Nelson, 2020), criticized the value of replication as arbiter of scientific legitimacy (Feest, 2019) and suggested that replications may be more common and failed replication more productive than critics claim (Guttinger, 2018; Guttinger and Love, 2019). In line with the attention to epistemic diversity in STS, scholars have criticized the ambitions of metascientists to push a narrow vision of replication across the sciences (Penders et al., 2019, 2020) and offered typologies to highlight why using replication to validate results becomes difficult, if not impossible, in some contexts (Leonelli, 2018). Others have discussed whether metascientists

have too narrowly interpreted the ‘crisis’, ignoring massive cultural and institutional changes to focus on more tractable issues of researcher bias (Flis, 2019) or poor statistical practice (Saltelli and Funtowicz, 2017).

This article presents the findings from a set of interviews with members of the Board of Reviewing Editors at *Science* and develops six theses which can put the discussion on new footing. The first two argue that much of the recent replication talk has been monopolized by a type of antagonistic replication we label ‘diagnostic’. By looking at how experimenters talk about replication as a routine part of laboratory practice, we show much replication activity is going unrecognized when the attention is limited to public tests of truth. We contrast the diagnostic motive for replication with an ‘integrative’ motive in which scientists replicate in order to adopt and extend rather than verify.

The second pair of theses highlight how the categories of ‘diagnostic’ and ‘integrative’ replication provide a framework to analyse how different sciences approach replication. Our respondents reveal that the vast majority of their replication activities are motivated by the desire to integrate. Yet, in some fields, integrative replication yields strong diagnostic evidence while in others it provides weak diagnostic evidence. We show that fields where routine experimental procedures are less standardized and outcomes are more uncertain tend to pursue more piecemeal and pragmatic replications, which provides weaker evidence regarding a finding’s veracity.

The final two theses argue that scientists who do less diagnostic replication are not simply abandoning the ideal of self-correction. Rather, our respondents offered a different, competing version of self-correction based upon an organic evolution of the field toward true findings rather than a formal process of explicit correction. This suggests that the activism currently transforming scientific practice is driven by a specific ideology of science rather than universally accepted principles of ‘good science’.

Replication, task uncertainty, and epistemic diversity

Beyond its epistemic value, replication has been of interest to sociologists as a social practice central to the organization of scientific communities. For Merton, the mere possibility of replication served a powerful social control function. Famously appealing to the ‘virtual absence of fraud in the annals of science’ (Merton, 1973: 276), he argued that replication kept scientists honest and careful and, thus, improved science as a whole. As Zuckerman (1977) later argued, the real power of replication is not in the replication itself, but in its anticipation. Given the risk of reputational damage, it is wiser to anticipate the threat of a failed replication and only share unimpeachable work. The idea that scientists internalize feelings of ongoing surveillance leads Shapin (1994: 413) to quip that ‘The modern place of knowledge [in Merton’s account] appears not as a gentleman’s drawing room but as a great Panopticon of Truth’. Mutual surveillance and critique are the basis of a ‘hidden hand’ mechanism that supposedly aligns the selfish motives of the individual scientist with the social good by rewarding good science and punishing non-replicable claims (Hull, 2001: 145).

While a central critique of metascientists is that the Mertonian vision has been insufficiently realized, STS scholars have undermined the central premise in the Mertonian account – that replication can, in fact, separate true from false claims and, thus, serve the

social control role the functionalist tradition (and, currently, metascientists) assign it. Rather than a ‘Supreme Court of the scientific system’ (p. 19) handing down clear verdicts of success or failure, Collins (1985) has shown that replications are often caught in an ‘experimenter’s regress’ in which scientists quibble about divergences of skill or equipment. Collins explains, ‘The “rule of replicability” provides a methodological prescription for scientists: “Replicate your observations or have them replicated!” But like any other rule, the rule of replicability does not contain the rules for its own application’ (Collins, 1991: 130).² Studies of contentious replication debates (e.g. Kennefick, 2000; Pinch, 1979; Travis, 1981) and the difficulties communicating and transferring tacit knowledge (Collins, 2001; Delamont and Atkinson, 2001; Doing, 2004) show how widespread these issues are.

Whitley (2000), in a comparative study of experimental practices across the sciences, has suggested that fields can be distinguished by their level of ‘technical task uncertainty’. In fields with greater task uncertainty, ‘technical procedures will be highly tacit, personal and fluid’ (p. 121). Such conditions are more likely to produce more ambiguous replications than fields where technical procedures are explicit, impersonal, and formalized.³ Variations in task uncertainty, a concept we elaborate on later, can result in different types of replication practice and interpretation.

STS research on replication, tacit knowledge, and epistemic diversity is necessary background for our forthcoming argument (Guttinger, 2018; Guttinger and Love, 2019; Leonelli, 2018; Penders et al., 2019, 2020; Pickersgill, 2014). We contribute a distinction to this tradition. Our hope is that it will help integrate these literatures in order to show how STS can contribute to the discussion currently being led by metascientists. Specifically, we argue that previous studies have failed to distinguish that replication has at least two, distinct motivations. *Diagnostic replication* is concerned with evaluating the truth value of a claim whereas *integrative replication* is concerned with incorporating findings from a study for one’s own purposes. Under certain conditions, these motives overlap. In others, however, they vary. Herein lies a key to understanding diverging replication practices across science and the logic behind competing ideals of self-correction in science.⁴

Methods

Rather than look at contentious replications, our focus is how scientists across fields replicate (or fail to replicate) research as a routine part of lab practice. Thus, instead of a case study which necessarily takes place within a specific epistemic context, we sought to engage with researchers from many diverse fields to get a more complete picture of the landscape of replication practice.

From September 2018 to January 2019, we contacted all 189 members of *Science’s* Board of Reviewing Editors (BORE). Ultimately, we conducted interviews with 60 members. 58 interviews were recorded and transcribed. One interview was conducted but, at the interviewee’s request, not recorded. One BORE member was interviewed over email. The in-person interviews ranged from 20 to 68 minutes with a mean of 46:33.

BORE members are the field experts whom *Science’s* managing editors rely upon to perform initial reviews of submissions and make recommendations about full peer review. We chose to interview them for three reasons. First, membership on the BORE is a useful proxy for expertise and professional reputation. Second, because *Science* is a generalist

journal, the BORE includes experts from every major field of natural science and the quantitative and experimental parts of social science. This gave us the opportunity to see how discussions about reproducibility were affecting a variety of scientific fields. Finally, since *Science* has been an active and vocal participant in these discussions, we thought that members of the BORE would be more aware of and sensitive to reproducibility.

Respondents were diverse both geographically and disciplinarily. Respondents hailed from 12 countries across five continents (although nearly all were from Europe [n=30] and the US [n=26]). Disciplinarily, respondents can be very roughly divided into life scientists (n=36), physical scientists (n=16), and social scientists (n=8).

Although this sample of elite scientists is diverse in many respects, the BORE is not a perfect representation of experimental science. Our respondents tended to be older, have higher status and be employed at wealthier institutions than the average scientist. These characteristics may incline them toward skepticism when faced with calls for reform. However, as we argue later, the harmony of their opinions with extant work in STS suggests that these are not simply the statements of elite scientists defending their privilege, but reflect unique aspects of their research conditions.

Interviews were structured around three broad areas: editorial work for *Science* and other outlets, conceptual issues of reproducibility, and how reproducibility was handled as a practical matter in their field. Regarding this last of these, we asked researchers how often they replicated research, how often such replications were successful, how they interpreted replication problems, and what they did to overcome issues when they arose. Thus, rather than focusing explicitly on contentious replication events, we asked respondents about the meaning of replication under routine circumstances. A coding scheme was developed inductively and all interviews were coded on ATLAS.ti software.

Interviews provide access to what people say and not what they do (Jerolmack and Khan, 2014). In potentially controversial areas, especially, interviews can be unreliable indicators of actual practice. However, our questions revolved around routine practice, not scandals. Moreover, the responses, even if they tended toward an idealized image of research (which, we will show, they clearly do not) differ in striking ways. In what follows, we present the statements without critical commentary, yet this should not be read as an endorsement of nor belief in the validity of the replication cultures they represent.

The integrative and diagnostic motives in replication

Integrative replication begins in trust. Diagnostic replication, on the other hand, is characteristic of Merton's 'organized skepticism' and, thus, is supposed to be agnostic about success or failure. Because the goal is simply to 'get it to work' rather than test the original finding, integrative replication is approached in a flexible and piecemeal manner. Conversely, the motivation behind diagnostic replication is explicitly to evaluate the original claim. A diagnostic replication is judged based upon its fidelity to the original study, not its outcome. Thus, replications pursued for integrative purposes can have dubious diagnostic value because of the protocol differences they admit.

Put simply, the goal of integrative replication is to reproduce the ends of research while being pragmatic about the means, whereas the goal of diagnostic replication is to faithfully reproduce the means while remaining agnostic about the ends.⁵

Success in an integrative replication attempt means reproducing results. Naturally, this often involves following a similar methodology, but the motive is pragmatic. If, for instance, a specific reagent is not available, another may be substituted in the hopes that it will work. If it does, then success has been achieved. Conversely, the success of a diagnostic replication is not found in its outcome but, rather, in its ability to claim to be an unimpeachable copy of the initial experiment.

Thesis 1: Strictly diagnostic replications are rare, but replications motivated by the desire to integrate are common.

The scientists we interviewed rarely conducted replications solely in order to test claims made in original studies. For instance, when we asked about hypothetically replicating a finding in his field, a biochemist replied, 'We'd never replicate it for the purpose of replicating it.'

Unless the original finding bore directly upon current research interests, replication was seen as an extravagance that labs could not afford. An immunologist explained, 'Investing a great deal of effort to reproduce what someone else has done just to prove that they were right is something we don't usually typically do.' Similarly, a neurobiologist told us, 'I'm very unlikely to have the person power to enable myself to spend time doing something like that.' And a biophysicist told us, 'Usually we just have enough time to work on what we are doing. So, it's hard to just ask a student to reproduce.'

In some cases, diagnostic replication was seen as a waste of time given their trust in their colleagues. A plant biologist said, 'We don't go into looking at these sorts of things with a view that they won't pan out. Our assumption is they will pan out.' A second plant biologist told us, 'I take what is published as true. I don't think I should double-check that it is right.' However, he then explained that replication is still occurring: 'Only when somehow this becomes relevant for my own research, at that moment I will start to recapitulate this experimentation.'

The lack of replications done for strictly diagnostic purposes did not reflect skepticism about or a disinterest in replication. However, replication was valued because it offered an opportunity to advance their own work, not because it allowed them to test the claims of others. As Firestein (2015: 151) writes, 'experiments get replicated because people from other labs use the published results and the methods in their own experiments'. A microbiologist told us, 'I tend to replicate it not because I want to replicate it, but if I want to base a research project on [it]'. A quantum physicist noted that he typically replicates because a novel finding 'frequently opens up new avenues of research, maybe new experimental capabilities.' Thus, rather than explicit checks for veracity, replication is conducted with an eye toward developing new capacities or pursuing new questions.

Because the goal of integrative replication is the development of one's own interests, it is often pursued in a selective and pragmatic manner, sometimes leading to 'microreplications' (Guttinger, 2018). As an immunologist explained, 'Normally you don't reproduce the same experiments. You just use the information to do something else.' A bioroboticist said, 'If it's an exciting approach or things like that then people will try to replicate and build on top of it.' However, because the focus of research is often originality, replications are often given some creative spin by scholars. He continued, 'At the

same time, you want novelty, so some people might explore other ways to do the same thing or doing it better. So, it's not very systematic.'

Thesis 2: Integrative replication attempts provide varying degrees of diagnostic evidence.

Members of the BORE reported that replication was common, yet its motivation was integrative and not diagnostic. Of course, these motivations are not wholly separate. Trying to integrate some process or technology provides a measure of evidence regarding its truth. Using a familiar metaphor, a genomics researcher explained, 'science is about building on other scientists' shoulders, right?' He went on to warn that if researchers publish irreproducible data, 'then people find out'. Along these lines, an economist said, 'important discoveries are those, of course, that create a platform on which others will build. And that's when they discover whether the platform itself is sound.' For a finding to become what Latour (1987) referred to as an 'obligatory passage point', other research groups must travel through it. They must adopt the technique, import the technology or base future studies on some purported relationship. If the original finding is wrong or is too difficult to recreate, it will languish – a dead end rather than a passage.

Several respondents made it clear that, although replications are common and important, they rarely try to faithfully recreate the original experiment. Instead, they prefer to pursue a piecemeal approach in which they try to reproduce the parts of the experiment that are most relevant to them. Yet, they still argued that this fragmentary replication provided some important information. For instance, a geneticist said, 'You rarely directly replicate somebody's work, but you frequently would do an experiment that would notice if somebody else's work was likely incorrect. Similarly, an immunologist explained that, although no lab will 'copy' studies, 'it's part of what they're studying so they're going to inevitably reproduce the findings. And, over time, it becomes clear what's reproducible and what's not'.

However, because the goal of integrative replication is not to explicitly test an original study, it raises important questions. What is their diagnostic power? To what degree can integrative replication – which, because of its pragmatic motive, is typically unsystematic – provide significant diagnostic information? This differs based on the relative degree of task uncertainty in different experimental systems.

The effect of task uncertainty on the interpretation of replications

Research on the material systems of scientific experimentation has often highlighted their unpredictability. For instance, Rheinberger (1997: 134) calls experimental systems 'generators of surprise'. Yet, he admits that these systems can, at times, become stabilized and transform into 'devices for testing, into standardized kits, into procedures for making replicas' (p. 80).⁶

These poles reflect what, following Whitley (2000), we call high and low task uncertainty. High task uncertainty is characteristic of experimental contexts in which significant variables are either unknown or uncontrollable and/or experimental techniques and technologies are either unstandardized or unstandardizable. In conditions of low task

uncertainty, on the other hand, variables are known and controlled and experimental techniques and technologies are standardized and predictable.

Whitley divides task uncertainty into ‘technical task uncertainty’ – which is high when experimental systems are significantly influenced by variables that are unknown, nonstandard or difficult to communicate (Devezer et al., 2020) – and ‘strategic task uncertainty’ – which is high when there is ‘uncertainty about intellectual priorities, the significance of research topics and preferred ways of tackling them, the likely reputational pay-off of different research strategies, and the relevance of task outcomes for collective intellectual goals’ (Whitley, 2000: 123). While there are both ontological and social dimensions, Whitley notes that this boundary is fluid. So, for instance, a field can purposely reduce technical task uncertainty by narrowly defining their subject matter and limiting methods considered appropriate for discovery. Because our respondents overwhelmingly interpreted uncertainty in technical terms, we will use the general term ‘task uncertainty’ to refer to situations where *respondents reported* that a lack of standardization, routinization, and formalization created difficulties in the communication and evaluation of research findings. We return to this potential importance of this distinction in the discussion.

The interpretation of replications, especially failed replications, is dependent upon the perceived degree of task uncertainty. Researchers try to integrate previous results, techniques, and technologies into their work. In experimental contexts with low task uncertainty, where standardization is expected to produce replicable outcomes, replication attempts are felt to be strong tests of original claims. In more uncertain contexts, where experiments are affected by a dizzying variety of variables both known and unknown, researchers are more reluctant to treat replications as definitive evidence for the veracity of the original study.

Thesis 3: Integrative replication provides stronger diagnostic evidence when task uncertainty is lower.

An integrative replication attempt provides powerful diagnostic evidence for a replicated study when task uncertainty is low. At one extreme, diagnostic power comes from the fact that replication is little more than recalculation. This is common in, for instance, some areas of macroeconomics where researchers share datasets. An economist explained, ‘When people look at large datasets that are publicly downloadable, that are available to all of the research teams, I think you find a genuinely large amount of true replication.’ Ironically, this sort of ongoing verification may actually appear to outsiders to reflect fewer replication attempts since these routine checks do not make it into the published record. The economist continued, ‘But, of course, people don’t bother writing me saying, “I completely replicated your results.” Because there’s no need to do that.’

In protein crystallography, researchers do not share the same datasets, but the structures are heavily constrained by a number of factors which leads researchers to believe that claims can be rigorously verified. A structural biologist argued his field had few reproducibility problems because there ‘are numerous opportunities to crosscheck whether the obtained results make sense’ using sequence and chemical data. Again, the power of these verification mechanisms in experimental systems with low task uncertainty actually creates a disincentive to replicate. When we asked a different structural

biologist if he felt what was published was sufficient to directly replicate findings in his field he responded, 'Yes. Yes. The thing is I really wouldn't say it was a productive enterprise because, provided that the statistics published the first time were good, I have zero expectation that I would find something different.' Rather than Rheinberger's 'generator of surprise', the expectation is that the experiment will produce an identical outcome making such a project a waste of time.

Task uncertainty is perceived to be lower when labs share material and technical culture. When techniques and technologies are standardized and shared across labs, there was an expectation among some respondents that results would replicate. A material scientist put it succinctly: 'When I have the same equipment that they do, then that's it. I should be able to find the same result.' A condensed matter physicist argued that 'Nobody can do something that somebody else cannot do. Ever. It's only a question of time. If a group does something, in six months, there should be another group doing it. There's no reason why not. Technologies are available. Techniques are standard.' He went on to say, 'If the result is really interesting, people want to reproduce it. And if they cannot reproduce then they consider it discarded. Then the question is whether it was because of incompetence or misbehavior.'

Because replication in contexts of low task uncertainty is interpreted as strong diagnostic evidence, the only question is about the motive behind the bad science. If a failed replication will lead your colleagues to brand you incompetent or a fraud, there is strong incentive to be conservative and exacting. These high reputational stakes explain why, when we asked how often he had been unable to reproduce findings in his field, a biochemist replied, 'That hasn't happened.' Under these conditions, even rare cases of fraud may be considered further evidence of the system working. When we asked a physicist about the case of Jan Hendrik Schön, who fabricated data on a series of publications in the early 2000s, he replied, 'The Schön thing. I can get really annoyed about it. So, it is often quoted as the example of big calamities occurring also in physics. But that's not really the story.' To him, this was less a story about sweeping replication problems in physics than it was about one brazen conman willing to trade short-term fame for his career: 'There was no way that he could escape it because you cannot survive when the whole world is trying to reproduce you.'

Amongst researchers in experimental cultures with low task uncertainty, there was an appreciation that it was a fortuitous situation that was not universal among experimental sciences. When we asked a systems biologist about reproducibility problems in his field, he replied, 'I think in our field it should be avoidable.' Yet, he admitted that the situation in his field was 'a bit different from other fields where the reproducibility issues can come from types of experiments which are really intrinsically hard to reproduce for different, diverse reasons'. Similarly, a structural biologist reflected, 'We're lucky we've chosen to work in this field because that kind of certainty appeals to us.'

Not all experimenters are so lucky. In fields where researchers perceive task uncertainty to be higher, replication is both more challenging and more difficult to evaluate.

Thesis 4: Integrative replication provides weaker diagnostic evidence when task uncertainty is higher.

When experimental technologies are idiosyncratic, when techniques require sophisticated embodied knowledge and/or when research objects are highly variable and hard to control, integration becomes more challenging and its ability to verify the original finding more open to interpretation. This is not to say that replication is not occurring. Integration is still being pursued, but it is more challenging, more piecemeal, and its diagnostic value is weaker.

Any dimension where labs differ can create interpretive problems. As a materials scientist explained how differences in equipment can create problems: 'Maybe the people that produced a result that you like so much, they did it with very sophisticated instrumentation, and yours is not up to their level.' Other times, differing levels of skill may be to blame. An immunologist conceded that he was often unwilling to use failed replications as evidence of problematic science because, 'it can be just because your postdoc working on it is not good enough'.

In many cases, however, researchers argued that reproducibility problems were simply unavoidable, given the unruly nature of their research objects. Several experimenters contrasted their research with experiments in 'less complex' or 'more controlled' fields. These were not meant as criticisms. Rather, they were offered as statements of fact, meant to appropriately calibrate expectations. After an endocrinologist detailed replication problems in his field, we asked to what he attributed them. He answered, 'I think the precision of the measurements in cell and molecular biology, and especially cell biology, is nothing like the precision of the measurements in physics. [I]n biology, even though it's a science just like physics is a science, there just aren't facts like that. They're our best approximations.' Making a similar argument, a cell biologist explained that he thought his field faced more problems with systemic complexity than a field like physics. He continued,

I hesitate to say that, because physics is complex, but I think many times we deal with systems that are difficult to control or difficult to know all the parameters that are relevant to the system. So, you're introducing more variability even though you might think you're doing the same experiment.

The uncertainty in the system means that experimenters often do not know why outcomes differ.

The unavoidable difficulty of working in certain areas demands a different interpretive framework for failed replications. A neuropathologist suggested that automatically interpreting failed replications as undermining original claims was 'naïve': 'Understand, when you are dealing with complex systems then experiments can come out completely different in different labs.' In stark contrast to the physicist who argued that failed replications in his field were evidence of 'incompetence or misbehavior', he argued, 'The reason is not because they are stupid or dishonest or sloppy. The reason is because the confounders are not known.'

Rather than a shameful secret, some researchers considered replication issues as evidence that they were working at the cutting edge of exciting possibilities. Rheinberger (1997: 28) argues that experimental systems are 'designed to give unknown answers to questions that the experimenters themselves are not yet able to clearly ask'. Under these conditions, failures can indicate that experimenters have a limited understanding of the 'parameter space of discovery' (Guttinger and Love, 2019). These are the conditions that

can produce the experimenters' regress, but they can also lead to a productive back-and-forth in which elements of the experiment that were hidden or tacit are made explicit and refined (Feest, 2016).

In this way, surprises and failures can be features rather than bugs. A plant biologist explained,

Lack of replication is not actually a problem. In biology, particularly, biological systems are messy and you do the experiment twice, the same people do exactly what they consider to be exactly the same experiment twice, and you get different results. ... Indeed, many exciting discoveries were made in precisely that way.

When even skilled researchers can do 'exactly the same experiment twice' only to get divergent outcomes, this weakens the power of diagnostic replication. Was the original study wrong? Or, are there potentially unknown variables for which we have not yet accounted?

Of course, defending one's research by appealing to unreported variables is often how scientific controversies get pulled into the experimenter's regress. Here, however, it was used in the opposite way. Rather than a strategy used to deepen a potential controversy, some respondents acknowledge these unreported variables to repair what might be seen as scandals of irreproducibility. For instance, a biomedical engineer dismissed the replication panic in her field. She argued that simply reading a report, following its methods section and expecting replication was a recipe for failure: 'I would not believe any of it I won't believe a lack of reproducibility unless there's a real concerted effort to communicate together and figure out where it might be coming from.' Similarly, a roboticist explained how replicating a robot can run into similar problems: 'Even if you're completely transparent about all the building blocks and really describe them in detail, putting all of everything together is still sometimes an art.'

Although this inability to definitively test claims may seem disappointing or, even, shocking to outsiders, members of fields may embrace 'epistemic modesty' (Pickersgill, 2016) in lights of these shortcomings. A marine biologist told us, 'I think there's an implicit understanding of that within the field, and therefore the interpretation of published work is pitched with appropriate sense around the uncertainty ... because of the lack of capacity to reproduce'.

These comments illustrate how the experimenter's regress can, ironically, become a category used to justify continued faith in the face of replication failures rather than just a pattern of controversy. They imply that replication *could* be achieved given the right amount of investment and signal a refusal to ascribe definitive meaning to failure. The greater the uncertainty in the experiment, the process of replication may involve talking on the phone, meeting in person, training next to someone, buying things, tinkering, etc. Slapdash attempts to integrate some finding can often be done quickly and cheaply. But it will yield little diagnostic information. However, the more fulsome the attempt, the costlier it becomes.

The price of diagnostic replication

The measure of a diagnostic replication attempt is its fidelity to the original experiment. The greater the difference between an original experiment and a replication, the less authority

can be claimed by the replication. When experiments have relatively low task uncertainty, this standard is easy to meet, even while pursuing one's own research interests. Conversely, in experiments with greater task uncertainty, fidelity to an original experiment may be more difficult to achieve and, more vexingly, it may not even be clear what fidelity entails.

When research objects are difficult to access, technologies are expensive and unevenly distributed or when experiments are heavily dependent on embodied skills, conducting a high-quality diagnostic replication can require special effort. This can create barriers to replication. A genetics researcher argued that replication was unlikely to deter fraud or sloppy science in his field because 'the experiments are so complicated and so convoluted' making it 'unusual for somebody to try to repeat precisely a body of work.' He continued, 'My lab published a year ago a body of work that involved, like, a quarter of a million dollars and five months' worth of very specific work. And if somebody wanted to go in and reproduce all of that work, it would cost them a whole boat-load of money.'

At times, the difficulty can be in acquiring a research object. An immunologist explained that it might take several months or a year to acquire a new knockout mouse and breed it before she could even begin to replicate a paper's claims. At other times, the experiment itself demands special skill. For instance, a cell biologist told us that, after he published his first major article which involved a complex experimental procedure, he told himself, 'I'm not sure anyone else is gonna do this experiment once they figure out how hard it is.' He explained,

I basically tried that experiment for about a year because, technically, it's incredibly challenging. And, it's not really the technology, it's the logistics. What we had to do is we had to look at one cell for about an hour, a living cell, take images. Then you have to take that cell off the microscope. They're on glass slides. There are thousands of cells on this glass slide. So, then you take this glass slide and we had to do a whole staining procedure and what have you and then find exactly the same cell again after we had done that staining. And you lose the cell, the staining doesn't work, what have you.

He noted that this was not unusual for cutting edge science. An article that debuts an important new technique might be the final product of multiple years of building and optimization: 'An inexperienced [replicator] who's all, like, "That's a real cool experiment. I've never done it but let me do it", they probably wouldn't do it right.'

When task uncertainty is high, faithful replications require special effort. They demand investments of time and money to gain the requisite skill and acquire the needed research objects and experimental technologies. And, still, failure – even uninterpretable failure – remains an option. In fields characterized by these types of uncertainties, researchers treat replication not primarily as a process of uncovering truth or falsity. Rather, replicators adopt an investment logic.

Thesis 5: When diagnostic replication requires special effort, experimentalists may embrace a logic of investment rather than a logic of truth.

The logic of diagnostic replication – what replication activists call for – is one of truth and falsity. An ideal replication is one in which the objects, instruments and techniques are similar enough to the original experiment to make its conclusion highly probative. When such similarity is hard, if not impossible, to achieve, a different logic emerges. A

cell biologist made this clear: ‘Often when we talk about reproducibility, we’re not really talking about an experiment working or not working. I think often, certainly in cell biology, it’s working better or worse ... It’s not black and white in that sense.’

When replication in the normal course of integration does not provide clear signals of truth or falsity, researchers instead look for signs of robustness. As an endocrinologist outlined, ‘I’m skeptical of everything. So, if my work has to build on it, we have to replicate it in my lab at least *at some level*’ (italics added). Yet, even when members of his lab failed to replicate something, he was unwilling to pass judgment because he did not feel that his attempt at integration provided strong diagnostic evidence about the original study. ‘I still don’t know it didn’t work because they did something wrong. They put the wrong thing in the wrong test tube, and this and that.’ Given the difficulty of replicating experiments with high task uncertainty, he often would have his lab members try a second time (even six or seven times if the finding could be particularly useful for the lab). Yet, if this whole process yielded no successes, he remained unwilling to say the initial claim was incorrect because he was unwilling to invest the resources to do what would constitute a true diagnostic replication.

Rather than assessing truth, his ultimate concern was where to most profitably direct his lab’s resources, embracing an ‘investment’ logic (Shiffrin et al., 2018: 2638). He made this explicit when he noted that ‘about 50 percent of the time the basic finding doesn’t work robustly enough for us to be worth our following up’. When we asked him if he thought that number was unacceptably high, he explained,

It’s a tough one because I’m not saying 50 percent of the time we wouldn’t be able to reproduce it if we put all our effort into it. That might be 10 or 20 percent, I’m not really sure. But again, a lot of the times, we’re just trying to figure out how to move forward. ... So, I wouldn’t say it’s incorrect at that point. I would just say with the effort we put into it, we’re not seeing evidence for this and it’s not that important to us, so let’s let it go.

Although he remained unwilling to pass judgment on the original study, failed replications gave him reason to direct his efforts elsewhere. Thus, even when outcomes are not black or white, failures are still informative. They told him that, regardless of the ultimate truth of the original claim, repeating it in his lab would require too much investment, given that other, more immediately profitable pathways were available.

This logic of investment was a common refrain. Given the options available, researchers preferred a tangibly ‘doable’ project to stubbornly trudging down a potential dead end. A molecular geneticist said, ‘I guess, what we’ve done more effectively than anything is that, if it looked like it wasn’t gonna work out, we dropped it real quickly. Rather than keep beating a dead horse if you know what I mean.’ A plant biologist made a similar point: ‘Most of us have a backlog of things to write up at the best of times; you’re going to prioritize things that you think are going to move things forward the most.’

Embracing a logic of investment produces a potential conundrum. Investing wisely, which often means cutting one’s losses rather than continue along a costly path with uncertain returns, means that researchers only interested in replicating research for their own interests are not doing the diagnostic work to actually see if claims are correct or not.

Metascientists have argued that such slapdash, unpublicized replications produce a poor culture of self-correction. Our respondents offered an alternative version of self-correction.

Modes of self-correction in science

When experimental systems have more task uncertainty, replication failures can be attributed to differences between labs – known, unknown and some, perhaps, unknowable. These differences can lead researchers to be reluctant in making diagnostic claims about their replication attempts. Professional courtesy means that experimentalists will often give the benefit of the doubt to their colleagues rather than assume sloppiness or bad faith.

As metascientists would likely point out, the decoupling of the integrative and diagnostic aspects of replication opens a space for irreproducible work to fester. For instance, a medical researcher told us that, during a postdoc working on the genetic aspects of cancer, he was getting scooped at every turn by a researcher at a different university: ‘I was incredibly frustrated, because I was thinking, “God this guy is so prolific, and he’s so famous, and he’s got golden hands.” I thought I was incompetent.’ Except the golden hands belonged to a fraud who had been falsifying his data and using this professional courtesy to elude detection.

It was clear that this situation was frustrating for some. After describing how his field of materials science was awash in claims of new materials with extraordinary properties, a researcher lamented that those claims often would not be backed up over time. He worried that researchers were exploiting replication difficulties to make grandiose claims and then rationalize away failed replications saying, “‘Oh yes, we were lucky. We had the magic batch, and we got these extraordinary properties, but we cannot reproduce because we cannot get a good batch without impurities today.” You cannot say whether they are true or wrong in the end.’ Similarly, an immunologist noted that, in her lab, failing to replicate ‘definitely happens a lot’. She continued, ‘People will publish some paper linking a gene to something, showing a phenotype in a mouse model. When you try to reproduce it yourself, it’s not as robust or sometimes it’s not there at all. There’s a lot of hand-waving that happens, I think particularly in the immunology world, about the role of the microbiomes.’ Although she acknowledged that microbiota were important and likely to contribute to some replication issues, she worried that too much was being attributed to them: ‘That’s definitely the easy one to throw your hands up at, but to me, that’s just not satisfying enough. You start to really question. ... Sometimes, it just seems too good to be true.’

It is this exact problem that has motivated metascientists to encourage forms of replication targeted to evaluate claims. In areas where integration provides only weak diagnostic evidence, researchers are no longer performing the work of policing themselves. Thus, metascientists argue that there needs to be more explicitly diagnostic replication. Yet, some members of the BORE suggested an alternative mode of self-correction in which explicitly diagnostic replication plays little role.

Thesis 6: When diagnostic replication is difficult and ambiguous, researchers may prefer an organic mode of self-correction.

None of the BORE members suggested that diagnostic replications were useless or should be avoided categorically. Some even engaged in such direct replications when they felt the stakes were high. Yet, the investment of time and money to conduct a highly faithful (and, thus, diagnostically potent) replication attempt, combined with its uncertain outcome, discouraged many researchers from pursuing this path.

The tension was illustrated by a parasitologist who was highly suspicious of an article that had been published that bore directly on his main research area. He was ‘absolutely stunned that it got into a journal like it did’ because ‘it didn’t make sense biologically’. Moreover, he was already suspicious of the first author on the paper for previous claims which he found dubious. Rather than rely on the system to correct itself, he decided to conduct a replication for the purposes of diagnosis: ‘We decided to try and repeat it, because it impacted the direction of our research and, if it was right, then we needed to change what we were doing.’

His team attempted a diagnostic replication, ‘directly reproducing what they had done’. In the end, his skepticism was supported, and the findings failed to replicate. They eventually published their work, but it took two full years and, after all their work, the original article was neither retracted nor was a correction issued. Noting both the cost and professional strife, the parasitologist lamented, ‘I probably wouldn’t do it that way again.’ Were the situation to recur, he suggested he ‘would go the usual way of self-correcting science, where you just keep going and prove your hypothesis using other experiments.’

Contrasting his lab’s diagnostic replication attempt with ‘self-correcting science’ may seem odd at first blush. After all, in the metascientific argument, diagnostic replications are the central mechanism in science’s self-correction. What is clear, however, from this and many other responses is that there are two, opposing conceptualizations of ‘self-correction’. We can label these ‘formal’ and ‘organic’.⁷

Formal self-correction occurs through the published literature. Its outcome is some change to the original study that either emends or retracts it. Replications of high diagnostic value are invaluable for formal self-correction in science. Formal replication attempts can be time-consuming, expensive and can create professional animosities. Yet, advocates argue that diagnostic replications are necessary to prevent the literature from being polluted with false positives which lead researchers to pursue doomed projects and, thus, wastes resources. Organic self-correction, on the other hand, happens largely through the unpublished backchannels of a field. Professional networks spread information about the relative reliability of claims gleaned during attempts to integrate them.

Formal self-correction remembers wrongness; organic self-correction forgets that which is not useful. A geophysicist said, ‘If we have a result and that result is not really acceptable, it will just die out. People are not referencing this work anymore and they will just move away’. An immunologist made the same point. Rather than need to be called out, an irreproducible finding is ‘forgotten by the community. But you don’t take the time to say it is wrong, you can’t reproduce it. You just do something else’.

The appeal to organic self-correction was common from experimentalists who worried about replication efforts redirecting resources away from the cutting-edge of the field. Thus, even after detailing replication problems in her field, an immunologist chafed at the idea of performing more diagnostic replications: ‘I have better ways to use my time. I would rather figure out what’s really going on and then include studies that might

argue against a previous publication, but not just publish a paper, send a response to a big paper just arguing that it's bogus'. While some may agree that such a reinvestment is worthy, others preferred a strategy in which the best science wins out by its ability to attract the most attention rather than through contentious, public trials.

Of course, metascientists encouraging a greater use of diagnostic replication have argued that organic self-correction has proven too weak to effectively police science. While scientists may interpret diagnostic replications as impolite or bullying, metascientists view this defensiveness as the reaction of an experimental culture that lacks the proper organized skepticism. Although we sympathize with this interpretation, it is important to understand that there are reasonable arguments that organic self-correction may be a better system under certain conditions. We turn to those arguments in the conclusion.

Discussion

Discussions about the role and nature of replication in science have ignored a key distinction between its diagnostic and integrative motives. This distinction has an important effect on how experimental communities approach and interpret replications. Where communities perceive a lack of standardized research objects, skills and instruments, direct replications may involve more of an investment than when replications can be conducted on standardized, at-hand equipment and research objects. Thus, researchers may 'test the waters' with a piecemeal replication and, if that fails, choose to move on to more profitable avenues rather than doggedly test a claim. The cost and ambiguity dissuade some scientists from investing in diagnostic replication as a mechanism for formal self-correction. Instead, they pursue an organic theory that 'the best science will win out'.

In detailing this view, we aim neither to uncritically accept nor valorize this image of scientific correction. However, the fact that accounts differ in predictable ways with task uncertainty suggests that there are real differences. Moreover, this alternative model of self-correction is not one that would be presented if the goal were to paint a rosy picture. Scholars embracing this view are essentially admitting that bad science may go uncorrected in the official record. Although they argue that knowledge of a study's robustness or fragility will still travel through backchannel networks, any form of self-correction that relies on private communication presents clear problems for a scientific system increasingly fragmented and global. Thus, this is neither an idealized nor unbelievable account. Rather, it is a plausible account of self-correction put forth that differs in key respects from the formal vision of the metascientists. Moreover, it is an account that in accord with research in STS that has emphasized the roles of tacit knowledge, interpretation, trust, and social negotiation in research.

In sum, we offer six theses which illustrate how the analytic distinction between diagnostic and integrative replication can help explain significant differences in how fields approach replication. To reiterate, the six theses are:

1. Strictly diagnostic replications are rare, but replications motivated by the desire to integrate are common.
2. Integrative replication attempts provide varying degrees of diagnostic evidence.

3. Integrative replication provides stronger diagnostic evidence when task uncertainty is lower.
4. Integrative replication provides weaker diagnostic evidence when task uncertainty is higher.
5. When diagnostic replication requires special effort, experimentalists often embrace a logic of investment rather than a logic of truth.
6. When diagnostic replication is difficult and ambiguous, researchers may prefer an organic mode of self-correction.

In the final two sections, we briefly outline insights that emerge from our analysis. The first section details how our argument reorients the discussion about replication in science studies. The second suggests how this work can inform current discussions around replication policy.

Rethinking replication in STS

The seemingly incongruous roles that replication plays in Mertonian and constructivist theories are made comprehensible when understood under our framework. The Mertonian ideal in which replication plays a powerful diagnostic role is common only in fields with low task uncertainty. In these cases, reproducing findings for one's own use provides useful diagnostic evidence about an initial claim. Yet, when task uncertainty is perceived to be high, replication becomes more challenging, and failures more ambiguous. This has been the focus of constructivists who have underscored contentions around replication. Thus, rather than disagree, the two theories correctly diagnose the power of replication and its limitations under different conditions.

Metascientists have suggested that the problems that lead to non-replicable findings tend to mirror the 'hierarchy of science' with physical science having the fewest problems, biological science a middling amount, and the human/social sciences the most (Fanelli, 2010; Fanelli and Ioannidis, 2013). Although our discussion does not directly address this, there are telling mismatches between fields that demand more study. For instance, despite falling at the intersection of physics and chemistry, the materials scientists admitted replication issues at odds with the positions of those fields. Conversely, despite being a biological science, structural biologists felt like replication problems were minimal. However, these apparent incongruities make sense when looking at their practices and community appraisals of uncertainty rather than simply where they fall in clumsy academic divisions. Doing so highlights that materials scientists operate in a field where studies can be published based on single (and, thus, unstandardized) batches. Similarly, constraints from multiple, standardized streams of evidence reduces uncertainty for the structural biologists.

These mismatches highlight the complex interrelation between the ontological and sociological aspects of task uncertainty. On one hand, the objects of study offer unique affordances which can lower or increase task uncertainty. Ion channels can be measured more accurately than human emotion. Planetary movement is more predictable than economic trends. On the other hand, the professional community plays an inextricable role in defining objects and determining standards. This raises the possibility that replicability problems may have a political solution. Those who exert control over a field can mandate

data, methods and analytic standards that reward replicable research (Fourcade et al., 2015; Pfeffer, 1993). However, the push for more technically exacting science can be done as the cost of research that is more exploratory, uncertain, and meaningful (Peterson, 2017). As activists and organizations seek to reduce replication problems through social change, the looming question is what is gained or lost by different strategies.

This raises a related question: under what conditions do replication failures constitute a ‘crisis’? Some of our respondents acknowledged replication problems in their fields while rejecting the moral panic propagated by metascientific activists. Meanwhile, these same activists would likely accuse such scientists of complacency in the face of ominous signs. The uneven success of metascientific activism in controlling the narrative and changing policies at journals, funding agencies, and institutions is an important topic at the intersection of STS, professions, and social movements.

Finally, STS scholars interested in how replication has become a rallying cry for epistemic activists need to better understand the outsized role of social psychology in this movement. Although metascience activism is seeking changes across the scientific landscape, many of the origins of the modern ‘crisis’ stem from social psychology. As such, activists have touted the field as a mix between the canary in the coal mine and a synecdoche for science itself. Although the level of agreement within the field should not be overstated,⁸ the interpretation that has dominated policy discussions and news accounts has been propagated by a small group of activist social psychologists.

Social psychology, however, faces unique challenges. Ethical limitations on human experimentation are compounded by the challenges of operationalizing the living, evolving concepts of everyday life (Derksen, 2017; Peterson, 2015). Although it is beyond the scope of this article, there are reasons to suggest that the current crisis represents the ‘psychologization’ of problems that have been around since the previous crisis in the 1960s and 70s (Faye, 2012; Morawski, 2020). Rather than confronting this task uncertainty, activists have interpreted the problem as psychological in nature – specifically, as the inevitable result of scientists’ inherent biases (Flis, 2019).

What needs to be scrutinized is the way the problems and solutions of social psychology have been extrapolated to the rest of the sciences. Although framed as a purely epistemic movement, reforms inevitably privilege some forms of knowledge and empower particular actors. As Leonelli (2018: 13) notes, ‘Generally, the emphasis on a narrow interpretation of reproducibility is linked with a devaluing of the role of expertise and embodied knowledge in data production, processing and assessment’. The very construction of a ‘replication crisis’ is the result of a specific group of psychologists, data scientists, and open science activists with particular interests (Feest, 2019; Peterson and Panofsky, 2020), and it is important to understand that reforms might cause damage to some types of epistemic communities (Penders et al., 2019; 2020).

Implications for science policy

STS scholars have made significant contributions toward understanding replication as a social practice yet, beyond occasional references to Merton, policy discussions largely ignore this literature. This article helps bridge this gap by highlighting how differences in replication

cultures are likely to produce different outcomes in regards to both the methods designed to diagnose replication problems in literatures and the policies aimed at solving them.

The diversity of replication practices has serious implications for policies designed to improve replication rates. By encouraging replications and making them easier to perform, metascientific activists hope that replication will move from the realm of a possible, but rarely actualized, deterrent to something with real teeth. Given our analysis, such policies would be especially effective in fields with low task uncertainty, where replication is interpreted to be most diagnostic. Yet, respondents in these fields were the least likely to endorse the need for such developments since the threat of a potent and unambiguous diagnostic replication is seen as sufficient to discourage bad behavior in many of these fields. Conversely, in fields with greater perceived task uncertainty – including much of the biomedical and behavioral research fields that most replication activism targets – replication initiatives are likely to have less impact because these replications are more susceptible to falling into the experimenters' regress.

Our analysis suggests care should be taken in the introduction of explicitly diagnostic replications in epistemic cultures where high task uncertainty has previously discouraged them. Under these conditions, diagnostic replication can create professional tensions. It is easy to understand why. Trust animates integrative replication. Skepticism, on the other hand, animates diagnostic replication. In experimental systems with low task uncertainty, diagnosis can happen as a natural byproduct of integration. When there is greater perceived uncertainty, experimenters must *go out of their way* to perform high-value diagnostic replications. Resistance to replication initiatives often focus on how these efforts damage research cultures by increasing suspicion. Thus, legal scholar Cass Sunstein recently tweeted (and, quickly deleted), 'in another life, the replication police would be Stasi'. The very idea of 'replication police' only makes sense in cultures where the integrative and diagnostic motive diverge, thus creating the possibility a new social role. Where task uncertainty is low, the 'policing' is already woven into uncontroversial lab practice.

This highlights the tension that can arise between the two motives for replication. Researchers have an inherent motivation during integrative replication. The goal is to 'get it to work' in order to extend their own research capacities. When replication is undertaken purely for diagnostic reasons, the motivations are unclear. What would motivate researchers to stop their own research in order to explicitly test a finding that is not integral to their own projects? Rather than 'get it to work', scholars conducting purely diagnostic replication attempts may, in fact, have a perverse incentive to have it fail. At the very least, they may lack the motivation to do it correctly. This can create a culture of paranoia which, while in line with the abstract ideal of 'organized skepticism', reflects a mistrust that is actually quite unusual in the history of science (Shapin, 1994: 16–20).

Finally, a more technical issue that arises from this analysis has to do with the evidence that metascientists have been using to diagnose problems in experimental literature. Metascientists have argued that having failed studies languishing in file drawers means that we are getting an incomplete picture of the data in a field. Our respondents made clear that, in many cases, they choose not to share such failures because the data was never meant to be diagnostic. They were quick and sloppy attempts to try something out, not sober processes of verification, and what might constitute a publicly sharable output is not

clear. This raises significant and complex questions regarding what sorts of failures should count and which should not be included in metanalytic analyses.


Acknowledgements

The authors would like to thank audiences at UCLA, UCSD, and 4S and the reviewers for their many helpful comments and suggestions.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project was funded by NSF grant #1734683.

ORCID iD

David Peterson  <https://orcid.org/0000-0003-4033-2993>

Notes

1. See Makel et al. (2012) on replication rate in psychology. This intersects with the ‘file-drawer problem’ (Rosenthal, 1979) because, it is argued, that even when replications are, in fact, conducted, they tend to languish, unpublished in file-drawers (Spellman, 2012).
2. Similarly, Lynch (2002: 211) has shown that, even in the case of seemingly standardized scientific practices, ‘Protocols do not, and cannot, fully explicate the work of reproducing them under singular circumstances’. Yet, during public trials – whether the replication controversies that Collins details or the legal case that is Lynch’s example – reliance on tacit knowledge and expert judgment becomes a source of possible dispute that replications themselves cannot adjudicate.
3. A similar conclusion can be gleaned from Cetina’s (1999) comparative ethnography of high energy particle physicists and molecular biologists. Because the practice of molecular biology is highly embodied in a way that particle physics is not, it would be expected that replication would be more uncertain.
4. Previous accounts have not been blind to these different aspects of replication practice. For instance, Zuckerman, who is centrally interested in the social control aspects of diagnostic replication, notes, ‘In the process of research, even long accepted contributions periodically come under renewed scrutiny, not by design but as a by-product of using them for further work’ (1977: 93). To put it into language we use later, integrative replication can, at times, provide important diagnostic evidence. Similarly, among case studies of contentious, explicit (and, in our terminology, diagnostic) replication attempts in gravity wave physics and parapsychology, Collins (1985) discusses what we refer to as integrative replication in the context of trying to build a TEA laser. Again, although present, neither tradition makes this distinction explicit.
5. There is some overlap between the diagnostic/integrative and direct/conceptual distinctions in replication. However, the two pairs of concepts differ importantly. While the latter describes different *forms of replication*, the former describes different *motives for replication*. Thus, while diagnostic replications always aim to be direct replications, those motivated for integrative purposes can range from more direct to more conceptual (i.e., less faithful to the materials and protocols of the original). As we explain, the factors that go into the decision to replicate closely or engage in more ‘quick and dirty’ replications help explain the evolution of different cultures of replication and philosophies of scientific self-correction.
6. These are dynamic processes. At times, what begins as an uncertain task becomes more predictable as key variables are discovered and controlled, standards created, and technologies stabilized. This ‘black boxing’ is a common feature of technoscientific progress (Latour,

- 1987). Yet, uncertainty can return when systems get reopened as they become too constraining (Jordan and Lynch, 1998: 795) or as part of an epistemo-political project (Nelson, 2020).
7. Although they do not use the term ‘organic’ to describe it, Shiffrin et al. (2018) draw on an extended metaphor of science-as-a-sprouting-tree to explain how science continues to make progress in the face of apparent reproducibility problems. Some nubs become branches while others do not attract resources. This is very much in line with organic self-correction.
 8. The idea that psychology is facing a crisis of replication or poor research practice has been disputed (Fiedler and Schwarz, 2016; Gilbert et al., 2016). Yet, even among those who advocate for change, there are questions regarding whether replication should be the focus of reforms. Some have, for instance, suggested that developing formal models or elaborating theories would be more profitable for the field than replicating studies (Devezer et al., 2019; 2020; Guest and Martin, 2020; van Rooij and Baggio, 2020).

References

- Baker M (2016) 1,500 scientists lift the lid on reproducibility. Available at: <https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970> (accessed 12 August 2019).
- Begley S (2015) In cancer science, many ‘discoveries’ don’t hold up. Available at: <https://www.reuters.com/article/us-science-cancer/in-cancer-science-many-discoveries-dont-hold-up-idUSBRE82R12P20120328> (accessed 12 August 2019).
- Cetina KK (1999) *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge, MA: Harvard University Press.
- Collins HM (1985) *Changing Order: Replication and Induction in Scientific Practice*. Chicago, IL: University of Chicago Press.
- Collins HM (1991) The meaning of replication and the sciences of economics. *History of Political Economy* 23(1): 123–142.
- Collins HM (2001) Tacit knowledge, trust, and the Q of sapphire. *Social Studies of Science* 31(1): 71–85.
- Delamont A and Atkinson P (2001) Doctoring uncertainty: Mastering craft knowledge. *Social Studies of Science* 31(1): 87–107.
- Derksen M (2017) *Histories of Human Engineering: Tact and Technology*. Cambridge: Cambridge University Press.
- Devezer B, Nardin LG, Baumgaertner B, et al. (2019) Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLoS One* 14(5): e0216125.
- Devezer B, Navarro DJ, Vandekerckhove J, et al. (2020) The case for formal methodology in scientific reform. *bioRxiv*. DOI: 10.1101/2020.04.26.048306.
- Doing P (2004) ‘Lab hands’ and the ‘Scarlet O’: Epistemic politics and (scientific) labor. *Social Studies of Science* 34(3): 299–323.
- Fanelli D (2010) ‘Positive’ results increase down the hierarchy of the sciences. *PLoS One* 5(4): e10068.
- Fanelli D and Ioannidis JPA (2013) US studies may overestimate effect sizes in softer research. *Proceedings of the National Academy of Sciences* 110(37): 15031–15036.
- Faye C (2012) American social psychology: Examining the contours of the 1970s crisis. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 43(2): 514–521.
- Feest U (2016) The experimenters’ regress reconsidered: Replication, tacit knowledge, and the dynamics of knowledge generation. *Studies in the History and Philosophy of Science Part A* 58: 34–45.
- Feest U (2019) Why replication is overrated. *Philosophy of Science* 86(5): 895–905.

- Fiedler K and Schwarz N (2016) Questionable research practices revisited. *Social Psychological and Personality Science* 7(1): 45–52.
- Firestein S (2015) *Failure: Why Science is So Successful*. Oxford: Oxford University Press.
- Flis I (2019) Psychologists psychologizing scientific psychology: An epistemological reading of the replication crisis. *Theory and Psychology* 29(2): 158–181.
- Fourcade M, Ollion E and Algan Y (2015) The superiority of economics. *Journal of Economic Perspectives* 29(1): 89–114.
- Freese J and Peterson D (2018) The emergence of statistical objectivity: Changing ideas of epistemic vice and virtue in science. *Sociological Theory* 36(3): 289–313.
- Gilbert DT, King G, Pettigrew S, et al. (2016) Comment on ‘estimating the reproducibility of psychological science’. *Science* 351(6277): 1037.
- Guest O and Martin AE (2020) How computational modeling can force theory building in psychological science. *PsyArXiv*. DOI: 10.31234/osf.io/rybh9.
- Guttinger S (2018) Replications everywhere: Why the replication crisis might be less severe than it seems at first. *BioEssays* 40(7): 1800055.
- Guttinger S and Love AC (2019) Characterizing scientific failure: Putting the replication crisis in context. *EMBO Reports* 20(9): e48765.
- Hull DL (2001) *Science and Selection: Essays on the Biological Evolution and the Philosophy of Science*. Cambridge: Cambridge University Press.
- Ioannidis JPA (2012) Why science is not necessarily self-correcting. *Perspectives on Psychological Science* 7(6): 645–654.
- Jerolmack C and Kahn S (2014) Talk is cheap: Ethnography and the attitudinal fallacy. *Sociological Research and Methods* 43(2): 178–209.
- Jordan K and Lynch M (1998) The dissemination, standardization and routinization of a molecular biological technique. *Social Studies of Science* 28(5–6): 773–800.
- Kennefick D (2000) Star crushing: Theoretical practice and the theoreticians’ regress. *Social Studies of Science* 30(1): 5–40.
- Latour B (1987) *Science in Action: How to Follow Scientists and Engineers through Society*. Cambridge, MA: Harvard University Press.
- Leonelli S (2018) Rethinking reproducibility as a criterion for research quality. In: Fiorito L, Scheal S and Suprinyak CE (eds) *Curiosity, Imagination, and Surprise*. Bingley: Emerald Publishing, 129–146.
- Lynch M (2002) Protocols, practices, and the reproduction of technique in molecular biology. *British Journal of Sociology* 53(2): 203–220.
- Makel MC, Plucker JA and Hegarty B (2012) Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science* 7(6): 537–542.
- Merton RK (1973) *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago, IL: University of Chicago Press.
- Morawski J (2020) Psychologists’ psychologies of psychologists in a time of crisis. *History of Psychology* 23(2): 176–198.
- Nelson NC (2020) The methodologists: A unique category of scientific actors. *Engaging Science, Technology, and Society* 6: 20–33.
- Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science* 349(6251): aac4716.
- Penders B, De Rijcke S and Holbrook JB (2020) Science’s moral economy of repair: Replication and the circulation of reference. *Accountability in Research* 27(2): 107–113.
- Penders B, Holbrook JB and De Rijcke S (2019) Rinse and repeat: Understanding the value of replication across different ways of knowing. *Publications* 7(3): 52.

- Peterson D (2015) All that is solid: Bench-building at the frontiers of two experimental sciences. *American Sociological Review* 80(6): 1201–1225.
- Peterson D (2017) The depth of fields: Managing focus in the epistemic subcultures of mind and brain science. *Social Studies of Science* 47(1): 53–74.
- Peterson D and Panofsky A (2020) Metascience as a scientific social movement. *SocArXiv*. DOI: 10.31235/osf.io/4dsqa.
- Pfeffer J (1993) Barriers to the advance of organizational science: Paradigm development as a dependent variable. *The Academy of Management Review* 18(4): 599–620.
- Pickersgill M (2014) The endurance of uncertainty: Antisociality and ontological anarchy in British psychiatry, 1950–2010. *Science in Context* 27(1): 143–175.
- Pickersgill M (2016) Epistemic modesty, ostentatiousness and the uncertainties of epigenetics: On the knowledge machinery of (social) science. *The Sociological Review* 64(1): 186–202.
- Pinch TJ (1979) Normal explanations of the paranormal: The demarcation problem and fraud in parapsychology. *Social Studies of Science* 9(3): 329–348.
- Popper K (1934/2005) *The Logic of Scientific Discovery*. Abington, PA: Routledge.
- Porter TM (1995) *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press.
- Rheinberger HJ (1997) *Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube*. Redwood City, CA: Stanford University Press.
- Rosenthal R (1979) The file drawer problem and tolerance for null results. *Psychological Bulletin* 86(3): 638.
- Saltelli A and Funtowicz S (2017) What is science's crisis really about? *Futures* 91: 5–11.
- Science News (2015) Runners-up. *Science* 350(6267): 1458–1463.
- Shapin S (1994) *A Social History of Truth: Civility and Science in Seventeenth-Century England*. Chicago, IL: University of Chicago Press.
- Shiffrin RM, Börner K and Stigler SM (2018) Scientific progress despite irreproducibility: A seeming paradox. *Proceedings of the National Academy of Sciences* 115(11): 2632–2639.
- Spellman BA (2012) Introduction to the special section: Data, data, everywhere ... especially in my file drawer. *Perspectives on Psychological Science* 7(1): 58–59.
- Travis G and David L (1981) Replicating replication? Aspects of the social construction of learning in planarian worms. *Social Studies of Science* 11(1): 11–32.
- Van Rooij I and Baggio G (2020) Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *PsyArXiv*. DOI: 10.1177/1745691620970604.
- Whitley R (2000) *The Intellectual and Social Organization of the Sciences*. Oxford: Oxford University Press.
- Zuckerman H (1977) Deviant behavior and social control in science. In: Sagarin E (ed.) *Deviance and Social Change*. Beverly Hills, CA: Sage, 87–138.

Author biographies

David Peterson is a postdoctoral researcher in the UCLA sociology department. He is interested in the sciences of the mind and brain, the emergence of metascience, and the application of artificial intelligence to scientific research.

Aaron Panofsky is an Associate Professor in the Institute for Society and Genetics, Public Policy, and Sociology. He is a sociologist of science, knowledge, and culture with a special interest on the history, intellectual organization, and social implications of genetics.