European Journal for Philosophy of Science (2025) 15:5
https://doi.org/10.1007/s13194-024-00629-x

PAPER IN GENERAL PHILOSOPHY OF SCIENCE

# The replication crisis is less of a "crisis" in Lakatos' philosophy of science than it is in Popper's

**Mark Rubin**[1]

Received: 27 December 2023 / Accepted: 12 December 2024 / Published online: 9 January 2025


## Abstract

Popper's (1983, 2002) philosophy of science has enjoyed something of a renaissance in the wake of the replication crisis, offering a philosophical basis for the ensuing science reform movement. However, adherence to Popper's approach may also be at least partly responsible for the sense of "crisis" that has developed following multiple unexpected replication failures. In this article, I contrast Popper's approach with that of Lakatos (1978) as well as with a related but problematic approach called *naïve methodological falsificationism* (NMF; Lakatos, 1978). The Popperian approach is powerful because it is based on logical refutations, but its theories are noncausal and, therefore, potentially lacking in scientific value. In contrast, the Lakatosian approach considers causal theories, but it concedes that these theories are not logically refutable. Finally, NMF represents a hybrid approach that subjects Lakatosian causal theories to Popperian logical refutations. However, its tactic of temporarily accepting a ceteris paribus clause during theory testing may be viewed as scientifically inappropriate, epistemically inconsistent, and "completely redundant" (Lakatos, 1978, p. 40). I conclude that the replication "crisis" makes the most sense in the context of the Popperian and NMF approaches because it is only in these two approaches that the failure to replicate a previously corroborated theory represents a logical refutation of that theory. In contrast, such replication failures are less problematic in the Lakatosian approach because they do not logically refute theories. Indeed, in the Lakatosian approach, replication failures can be temporarily ignored or used to motivate theory development.

**Keywords** Metascience · Popper · Philosophy of science · Lakatos · Replication crisis · Theory testing

The replication crisis occurred because replication rates were lower than "expected or desired" (Nosek et al., 2022, p. 724; see also Munafò et al., 2017, p. 1; Open

✉ Mark Rubin
Mark.Rubin@durham.ac.uk

[1] Durham University, Durham, United Kingdom

Science Collaboration, 2015, p. 7), although it remains unclear what an "acceptable" replication rate should be (Rubin, 2023b, p. 4). A seminal attempt to replicate 100 psychology studies kick-started the crisis by finding that only 39% of effects were rated as replicable (Open Science Collaboration, 2015). Subsequent work has raised concerns about a replication crisis in other disciplines including, for example, preclinical cancer biology (Errington et al., 2021), economics (Chang & Li, 2022), and experimental social and behavioural science (Camerer et al., 2018).

Researchers have tended to focus on false positive results in original studies as the primary cause of the replication crisis. The assumption is that questionable research practices have inflated the "actual" false positive rate above the nominal level, leading to a surprising number of replication failures (e.g., Munafò et al., 2017).[1] In contrast, philosophers of science have focused less on poor research *practices* and more on general research *principles*. For example, Bird (2021) argued that the surprising number of replication failures is not caused by a larger than expected number of false positive results but by a larger than expected number of false hypotheses. Maziarz (2024) argued that ostensibly conflicting results may represent valid answers to different research questions rather than false positive results in one or other study. Finally, Rubin (2023b), Feest (2019), and to some extent Fletcher (2021, p. 8) have argued that the centrality of replications in science has been overplayed (cf. Sikorski & Andreoletti, 2023).

The replication crisis has also highlighted the distinction between *direct* (close) replications and *conceptual* replications (e.g., Crandall & Sherman, 2016; Derksen & Morawski, 2022; Feest, 2019; Nosek & Errington, 2020; Rubin, 2020; Sikorski & Andreoletti, 2023; Zwaan et al., 2018). Direct replications attempt to test the same hypothesis using the same conditions, methods, and analyses as in the original study, whereas conceptual replications test the same hypothesis using different conditions, methods, and/or analyses. Key issues here have been (a) how to define "same" and "different," (b) which type of replication has more scientific value, and (c) whether the distinction between "direct" and "conceptual" is useful. For example, Nosek and Errington (2020) described the distinction as "at least irrelevant and possibly counterproductive for understanding replication and its role in advancing knowledge" (p. 2; see also Derksen & Morawski, 2022; Feest, 2019; Rubin, 2020).

Popper's (1983, 2002) work has figured prominently throughout these discussions. His approach is regarded as "useful for both understanding and remediating the replication crisis" (O'Donohue, 2021, p. 236), and the ensuing science reform program has been described as "distinctly Popperian" (Derksen, 2019, p. 460; see also Flis, 2019). Certainly, many aspects of Popper's work are useful in science in general and in relation to the replication crisis and science reform in particular. However, concerns about multiple unexpected replication failures may also be more relevant in the Popperian approach than in other approaches. From this perspective, adherence to the Popperian approach may have accentuated the sense of a replication "crisis." Accordingly, it is worth considering how other philosophies of science might characterise multiple unexpected replication failures.

The purpose of this article is to undertake a general philosophical assessment of the impact of replication failures in the context of three approaches in the philosophy of science: Popper's (1983, 2002) approach, Lakatos' (1978) approach, and a related but

problematic approach called *naïve methodological falsificationism* (Lakatos, 1978). I conclude that the replication crisis may be less of a "crisis" for Lakatosians.

Popper argued that theories are only scientific if they are *logically* falsifiable (e.g., Popper, 1983, pp. xix-xx), meaning that they have the potential to be falsified through a process of logical deduction. Lakatos (1978) disagreed. In a paper that Feyerabend (1975) described as "one of the most important achievements of twentieth-century philosophy," Lakatos (1970, 1978) proposed that "exactly the most admired scientific theories simply fail to forbid any observable state of affairs" (Lakatos, 1978, p. 16, italics omitted). Popper (1974b) responded that, "were the [Lakatosian] thesis true, then my philosophy of science would not only be completely mistaken, but would turn out to be completely uninteresting" (p. 1005).

To understand the reasons for this disagreement, I begin by considering the different ways in which Popper and Lakatos conceptualized scientific theories, focusing in particular on the concept of causality. For Popper, scientific theories are noncausal universal statements (e.g., "all swans are white"), whereas for Lakatos scientific theories specify causal connections (e.g., "swanness causes whiteness").

I then discuss the implications of these different conceptualisations for the logical falsifiability of scientific theories. Both Popper and Lakatos agreed that causal theories are not logically refutable. Hence, for Popper, scientific theories must be noncausal in order to be logically refutable, whereas for Lakatos, scientific theories must be causal and, therefore, not logically refutable.

I also discuss a third position called naïve methodological falsificationism (NMF). This approach assumes that causal theories can be made to be logically refutable by tentatively accepting a ceteris paribus clause that states that no other causally relevant factor is at work during the testing of the theory. I argue that this approach is problematic because accepting a ceteris paribus clause during theory testing is scientifically inappropriate, epistemically inconsistent, and "completely redundant" (Lakatos, 1978, p. 40).

I then consider the extent to which the replication crisis is a "crisis" within the Popperian, Lakatosian, and NMF approaches. I argue that replication failures are more impactful in the Popperian and NMF approaches than in the Lakatosian approach because it is only in these first two approaches that replication failures logically refute theories.

Finally, I consider some responses to the replication crisis. I argue that Popperian and NMF researchers may be reticent to adopt a "hidden moderator" explanation of replication failures because it precludes the logical refutation of extant theories. In contrast, Lakatosian researchers are more amenable to a hidden moderator explanation because they are concerned about the *development* of theories rather than their *logical refutation*. I close by discussing three alternative perspectives on the Lakatosian approach put forward by Zwaan et al. (2018b), Earp and Trafimow (2015), and Uygun-Tunç and Tunç (2023).

## 1 What is a scientific theory?

A concern in some areas of modern science is that researchers are focused on the collection of robust and reliable empirical effects rather than on testing substantive theory. This *naïve empiricism* (Forscher, 1963; Strong, 1991) makes inductive predictions about future effects based on reliable demonstrations of past effects. However, it does not focus on theoretical explanations of those effects. Consequently, although naïve empiricism may be able to tell us that an effect is likely to reoccur, it struggles to explain *why* the effect occurs or *when* and *where* it is unlikely to occur. Popper and Lakatos both rejected this naïve empiricist view of testing effects and instead focused on testing hypotheses and theories.

According to Popper, a *strictly universal statement* or law such as "all swans are white" can serve as a scientific hypothesis (Popper, 2002, p. 38). This hypothesis can be used to deduce a "negative" prediction, such as "there will be no non-white swans at this time and place." This "nonexistential proposition" is a "specialization of a universal law…to a particular space-time region"; Popper, 1974b, p. 998). However, there are no non-white swans in lots of space-time regions (Popper, 2002, p. 83), and "we cannot search the whole world in order to make sure that nothing exists which the law forbids" (Popper, 2002, p. 49). Consequently, the universal statement must be combined with (a) the *initial conditions* of a *specific* space-time region and (b) the *auxiliary hypotheses* of a test (which together constitute our *background knowledge*) in order to deduce a "potential falsifier" in that particular region. This potential falsifier takes the form of a "basic statement" (a singular existential statement that refers to a specific space-time region) that describes an intersubjectively observable event such as "there is a black swan in this location at this time" (see also Popper, 1974b, p. 997; Popper, 1983, p. xx; Popper, 2002, pp. 38, 83). Acceptance of this potential falsifier during hypothesis testing will then logically refute the hypothesis that "all swans are white." It is important to note that this logical refutation does not necessarily imply that we should *reject* and abandon the hypothesis (Popper, 1974b, p. 1009). Other non-logical matters must be taken into account before making this more substantive decision.

Lakatos (1978) did not disagree with the above reasoning. However, he argued that the hypothesis "all swans are white" is not a scientific theory. According to him, "a proposition might be said to be scientific only if it aims at expressing a causal connection" (pp. 18–19). For example, the proposition "swanness causes whiteness" (Lakatos, 1978, p. 19) represents a scientific theory because it expresses a causal connection (e.g., swan DNA causes white plumage; Karawita et al., 2023).

Contrary to Lakatos (1978), Popper (2002, p. 39) believed that the "principle of causality" should be excluded from science because it is a metaphysical concept. Hence, he would reject Lakatos' (1978, p. 19) proposal that the statement "swanness causes whiteness" represents a scientific theory. This is not to say that Popper ignored causal explanations. As he explained, "to give a causal explanation of an event means to deduce a statement which describes it, using as premises of the deduction one or more universal laws, together with certain singular

statements, the initial conditions" (Popper, 2002, p. 38, italics omitted). Popper gave an example in which (a) the hypothesis is "whenever a thread is loaded with a weight exceeding that which characterizes the tensile strength of the thread, then it will break"; (b) the two initial conditions are "the weight characteristic for this thread is 1*lb*," and "the weight put on this thread was 2*lb*s"; and (c) the (positive) prediction is "this thread will break" (Popper, 2002, p. 38). In this situation, the observation that "this thread broke" and the situation in which it occurred are the *explicandum* or "state of affairs to be explained," and the theory and its deduced prediction in relation to the initial conditions represent the independently testable explanation or *explicans* (Popper, 1983, p. 132). In addition, the initial conditions describe the "cause," and the prediction describes the "effect" (Popper, 2002, pp. 38–39).

Critically, however, and in contrast to Lakatos, Popperian hypotheses and theories are *noncausal* universal statements ("all swans are white") rather than causal connections ("swanness causes whiteness"). As in the above example, initial conditions and predictions may be described as "causes" and "effects" respectively. However, Popper (2002) preferred to avoid these terms, and he was clear that no "principle of causality" should be invoked (p. 39). As he explained, "I shall be content simply to exclude it [the principle of causality], as 'metaphysical', from the sphere of science" (Popper, 2002, p. 39; see also Popper, 2002, p. 48). Hence, what is logically refuted in a Popperian theory test is a *noncausal* universal statement rather than a causal connection.[2]

Lakatos (1978) was concerned that, in the absence of a metaphysical causal connection, a Popperian theory may be regarded as a "mere curiosity" or "oddity" without any obvious "scientific value" (pp. 18–19). *Why* are all swans white? By itself, a Popperian theory does not provide an explicit answer to this question. This situation was unsatisfactory for Lakatos, who argued that "science…must be demarcated from a curiosity shop where funny local – or cosmic – oddities are collected and displayed" (p. 18). Similarly, Pearce (1990) noted that Popper's "all swans are white" example "has little relevance to science since scientific theories are not generalizations of facts; rather, they involve an understanding of the underlying processes that *cause* certain facts to occur" (p. 47, my emphasis). Lakatosian theories provide this scientific relevance in the form of causal connections: "all swans are white *because* swanness causes whiteness." Theorists adopt this Lakatosian approach whenever they incorporate causal connections in their theorizing and modelling. For example, as Guest and Martin (2021) explained, "a theory is a scientific proposition – described by a collection of natural-language sentences, mathematics, logic, and figures – that introduces *causal relations* with the aim of describing, explaining, and/or predicting a set of phenomena (Lakatos, 1976…)" (p. 794, my emphasis).

In summary, for Popper, scientific theories must be logically falsifiable, whereas for Lakatos, they must be causal connections. In the Popperian approach, a logically falsifiable universal statement represents both a hypothesis and a theory (Popper, 2002, pp. 4, 37–38, 48; see also Hager, 2000, p. 5; Monnerjahn, 2019).[3] Hence, a logical refutation of a hypothesis is also a logical refutation of a scientific theory. In contrast, the Lakatosian approach provides a conceptual distinction between noncausal hypotheses and causal theories: A noncausal hypothesis (e.g., "all swans are white") is not

a scientific theory because it does not provide a causal connection (e.g., "swanness causes whiteness"). Consequently, for Lakatos, the logical refutation of a hypothesis does not necessarily imply the logical refutation of its associated theory (e.g., the refutation of "all swans are white" does not necessarily imply the refutation of "swanness causes whiteness"). Indeed, as I discuss next, Lakatos argued that causal theories are not logically refutable (for a similar view, see Putnam, 1991).

## 2 Lakatos' view that causal theories are not logically refutable

There are many causes in the universe, including some that may confound and counteract the particular cause that we are investigating in our study (Johansson, 1980). For example, a genetic factor may cause a swan to be black even though it remains true that swan DNA causes white plumage. The intervention of this counteracting cause would not logically refute our causal theory because it operates independently from our theorized cause. In addition, our theorised cause may be moderated by various factors. For example, swan DNA may only cause white plumage in some environments and not in others. Again, moderator factors do not refute the existence of putative causes; they merely limit their influence.

To acknowledge the potential impact of these confounding, counteracting, and moderating factors, we may attempt to delineate them within an exclusive *ceteris paribus* clause which states that various specified and unspecified causally relevant factors do not affect our results during theory testing (see also Putnam, 1991, p. 137; Trafimow & Fiedler, 2024). However, this clause may be incorrect because other relevant factors may, in fact, affect our test results. Hence, we need to acknowledge that a test of a causal theory is also a test of a fallible ceteris paribus clause. For example, we don't just test the theory that "swanness causes whiteness"; we test a conjunction of this theory and a ceteris paribus clause: "swanness causes whiteness *provided that no other relevant factor is at work*" (Lakatos, 1978, pp. 17–18). The observation of a black swan may then refute this proposition because either (a) swanness *does not* cause whiteness or (b) swanness *does* cause whiteness but some other relevant factor has intervened to produce a black swan. Hence, Lakatos argued that, although a black swan can logically refute the noncausal hypothesis that "all swans are white," it cannot logically refute the causal theory that "swanness causes whiteness" because it may instead refute the fallible ceteris paribus clause that "no other relevant factor is at work" (see also Putnam, 1991, p. 127).

Note that Lakatos' concern here was only about testing *causal theories*. He agreed with Popper's approach of testing *noncausal hypotheses*. In particular, he agreed with Popper that tentatively and temporarily accepting certain initial conditions and (noncausal) auxiliary hypotheses as unproblematic background knowledge during hypothesis testing "cannot be avoided" (Lakatos, 1978, pp. 42–43). However, as discussed above, he argued that tests of noncausal hypotheses per se provide an inadequate approach to science, and that tests of associated causal theories are not logically refutable.

Lakatos also accepted that it is possible for us to become more confident in our tests of causal theories by subjecting the ceteris paribus clause to severe tests.

> "How can one test a *ceteris paribus* clause severely? By assuming that there are other influencing factors, by specifying such factors, and by testing these specific assumptions. If many of them are refuted, the *ceteris paribus* clause will be regarded as well-corroborated" (Lakatos, 1978, p. 26, italics in original; for a similar approach, see Uygun-Tunç & Tunç, 2023).

Nonetheless, like Popper, he maintained that "the decision to 'accept' a ceteris paribus clause is a very risky one because of the grave consequences it implies" vis-à-vis the premature falsification of a causal theory (Lakatos, 1978, p. 26; see also Popper, 1974b, Footnote 75, pp. 1186–1187). Consequently, Lakatos' (1978, pp. 37–38) *sophisticated methodological falsificationism* does not tentatively accept the ceteris paribus clause during theory testing, and it cannot logically refute the theories it tests. Instead, in Lakatos' sophisticated approach, a theory is only regarded as being falsified when a new theory has been proposed that (a) accounts for the old evidence that previously supported the falsified theory, (b) makes new predictions that are improbable under the falsified theory, and (c) some of those new predictions are subsequently corroborated (Lakatos, 1978, p. 32). The new theory may be based on a modification of the falsified theory, rather than its outright rejection, where the modification involves the addition of one or more auxiliary hypotheses that qualify the scope of the falsified theory (Lakatos, 1978, p. 33).

## 3 Naïve methodological falsificationism

In summary, the Popperian approach is powerful because it is based on logical refutations, but its weakness is that its theories are noncausal and, therefore, potentially lacking in scientific value (a "mere curiosity"; Lakatos, 1978, p. 19). In contrast, the Lakatosian approach is powerful because it tests causal theories, but its weakness is that its theories are not logically refutable.

Lakatos (1978) noted the possibility of a third, problematic, approach to theory testing, which he described as *naïve methodological falsificationism* (NMF). From my perspective, NMF hybridizes the Popperian and Lakatosian approaches. It claims *both* the deductive power of Popper's logical refutations *and* the scientific relevance of Lakatos' causal theories. It does so by attempting to logically refute not only noncausal hypotheses ("all swans are white"), but also causal theories ("swanness causes whiteness"). Like the Popperian and Lakatosian approaches, the NMF approach influences appraisals of replication failures. Hence, I explain how it operates, and I consider its weaknesses.

According to Lakatos (1978), the NMF approach circumvents the logical problems associated with testing causal theories by temporarily and tentatively accepting the ceteris paribus clause that "no other relevant factor is at work" during theory testing. Consequently, the ceteris paribus clause is excluded from the test and the specific causal theory is left as the only remaining statement that can be logically refuted by an anomalous result. As Lakatos (1978) explained, "we may call an event described by a statement $A$ an '*anomaly* in relation to a theory $T$' if $A$ is a potential falsifier of the conjunction of $T$ and a *ceteris paribus* clause but it becomes a potential falsifier of $T$ itself after having decided to relegate the ceteris paribus clause

into 'unproblematic background knowledge'" (p. 26, italics in original). However, there are three related problems with the NMF approach of accepting ceteris paribus clauses as unproblematic (i.e., irrefutable) during theory testing.

### 3.1 Accepting ceteris paribus clauses is scientifically inappropriate

Accepting the ceteris paribus clause as temporarily "unproblematic" during a causal theory test changes the proposition under test from the logically irrefutable statement that "swanness causes whiteness *provided that* no other relevant factor is at work" to the logically refutable statement that "swanness causes whiteness *and* no other relevant factor is at work." One potential problem with this approach is that the proposition "no other relevant factor is at work" is unrealistic given the infinite range of potential factors to which it refers. For example, even if it is true that "swanness causes whiteness," it is unreasonable to accept that no other factor in the universe could cause a non-white swan. Consequently, we are left with a choice between (a) testing a proposition that cannot be logically refuted and (b) temporarily accepting a proposition that is unrealistic (for a discussion of a related dilemma, see Reutlinger et al., 2021, Sect. 3.4).

Of course, scientists often condition their tests on unrealistic, idealised, counterfactual models on the assumption that "all models are wrong, but some are useful" (Box & Draper, 1987, p. 424; see also Popper, 2002, p. 72). They may also introduce potential confounders during a process of de-idealisation in order to make their models more realistic. In both cases, however, NMF researchers test a proposition with an accepted ceteris paribus clause in order to allow the logical refutation of a causal theory. In contrast, Lakatosian researchers regard such a proposition as scientifically inappropriate because it prevents a consideration of whether causally relevant factors have influenced the test result. To be clear, Lakatosians may entertain the unrealistic assumption that no other relevant factor is operating when they test a causal theory. However, unlike, NMF researchers, they never make the methodological decision to tentatively and temporarily accept the ceteris paribus clause in order to force logical refutations of causal theories. They are *always* open to the possibility that other causes could have affected their test result and, consequently, their test result cannot logically refute theories.

### 3.2 Accepting ceteris paribus clauses is epistemically inconsistent

The NMF decision to accept the ceteris paribus clause as irrefutable, even on a tentative and temporary basis, is inconsistent with a scientist's epistemic obligation to specify their doubt and ignorance about the potential influence of other relevant factors in their investigations (e.g., Feynman, 1955; Firestein, 2012; Merton, 1987; Rubin, 2024). The ceteris paribus clause represents this doubt and ignorance. It is where scientists acknowledge both their "known unknowns" (what they know they don't know – their "specified ignorance"; Merton, 1987) and their "unknown unknowns" (what they don't know they don't know – their unspecified ignorance; Rubin, 2023a; Trafimow & Fielder, 2024, p. 8). Consequently, accepting a ceteris

paribus clause as "unproblematic" during a causal theory test flies in the face of scientific humility.

An NMF researcher's decision to temporarily accept the ceteris paribus clause is also inconsistent with (a) their future research activities and (b) their colleagues' ongoing research activities. How can a scientist "accept" that no other relevant causal factor is influential during their theory test and then go on to test the influence of some of those factors in their future work? Similarly, how can they accept a ceteris paribus clause as "unproblematic" when, all around them, their colleagues are busily investigating the influence of many of the factors it contains? As Meehl (1990, p. 111) explained,

> "for the ceteris paribus clause to be literally acceptable in most psychological research, one would have to make the absurd claim that whatever domain of theory is being studied (say, personality dynamics), all other domains have been thoroughly researched, and all the theoretical entities having causal efficacy on anything being manipulated or observed have been fully worked out! If that were the case, why are all those other psychologists still busy studying perception, learning, psycholinguistics, and so forth?"[4].

These various inconsistencies may be dismissed by arguing that researchers temporarily abandon the role of "scientist" and instead adopt the role of a "quality controller" who accepts the background knowledge of their test and automatically (logically) refutes products (theories) that do not meet the test's stated criteria (Rubin, 2020). However, this role-switching account does not resolve the problem of "epistemic inconsistency" (Rubin, 2020, p. 7): The logical refutation of a causal theory that is obtained in the role of quality controller becomes an illogical refutation when the quality controller returns to the role of scientist and begins, once again, to doubt the validity of the ceteris paribus clause.

### 3.3 Accepting ceteris paribus clauses is "completely redundant"

NMF researchers might argue that their acceptance of the ceteris paribus clause is only tentative and temporary, and that they will bring the clause back into question after their theory test. This position is consistent with Popper (1974b, p. 1009), who argued that the *logical refutation* of a theory does not necessarily imply that researchers should subsequently *reject* the theory in practice and stop working on it. But if this is the case, then what is the function of logical refutations during theory testing? Why should we temporarily and tentatively "accept" other relevant causes as being uninfluential during our test in order to force a logical refutation of a theory if we are only going to bring these other causes back into consideration when deciding whether to reject (stop working on) that theory? Instead, why not consider the logical refutation of noncausal *hypotheses* (e.g., "all swans are white") in the context of explanations provided by *both* fallible causal theories ("swanness causes whiteness") *and* alternative causal theories within the fallible ceteris paribus clause ("some other relevant factor is at work") and then come to a tentative conclusion in a process of inference to the best explanation (e.g., Haig, 2009; see also Popper, 1974a, pp. 15–16; Popper, 1983, p. 55; Popper, 2002, p. 438)? Lakatos had

a similar view, describing the NMF decision to temporarily accept the ceteris paribus clause during theory testing as "completely redundant" (Lakatos, 1978, p. 40) and the associated refutation as "utterly irrelevant" (Lakatos, 1968, p. 158).

### 3.4 Summary

The NMF approach represents a potentially powerful hybrid of the Popperian and Lakatosian approaches because it aims to logical refutate causal theories. However, the NMF decision to accept the ceteris paribus clause as temporarily irrefutable is problematic for three reasons. First, it is scientifically inappropriate because it prevents a consideration of other causally relevant factors as having a potential influence on the test result. Second, it results in an epistemic inconsistency because researchers accept propositions during theory testing that they subsequently doubt. Third, it is "completely redundant" (Lakatos, 1978, p. 40) because the logical refutation of a causal theory during testing does not necessarily imply its rejection in practice.

It is important to appreciate that Popper also rejected the NMF position. He explained that "the clause '*ceteris paribus*' (all other things being the same) is, of course, never satisfied in this world," and he agreed with Lakatos that the use of the clause is intended to imply that "the *relevant* circumstances should not change" (Popper, 1974b, Footnote 75, p. 1186). However, like Lakatos, he argued against the acceptance of such ceteris paribus clauses:

> "What is relevant or irrelevant is a matter of risky *conjecture* (and such conjectures will be the more interesting the more specific they are, and the more testable they render the original theory). I therefore suggest that *ceteris paribus* clauses should be avoided" (Popper, 1974b, Footnote 75, pp. 1186–1187, italics in original; see also Lakatos, 1978, p. 26).

Hence, Popper acknowledged scientists' ignorance about the relevant and irrelevant circumstances of their test and, on this basis, argued that "no *ceteris paribus* clause is necessary" (Popper, 1974b, Footnote 75, p. 1186). To be clear, in a Popperian theory test, we must tentatively and temporarily accept initial conditions and noncausal auxiliary hypotheses as unproblematic background knowledge. However, there is no need to tentatively accept the ceteris paribus clause that "no other relevant cause is at work" because we are not testing (metaphysical) causal theories.

## 4 How much of a "crisis" is the replication crisis?

The three approaches that I have considered each have their strengths and weaknesses. The Popperian approach can logically refute noncausal theories. However, the scientific value of these theories is unclear. The Lakatosian approach considers causal theories. However, in Lakatos' view, their causal nature makes them

logically irrefutable. Finally, the NMF approach aims for the best of both worlds by logically refuting causal theories. However, it does so by temporarily accepting ceteris paribus clauses, and this approach may be characterised as scientifically inappropriate, epistemically inconsistent, and completely redundant. Table 1 provides a summary of these three approaches.

What are the implications of the replication crisis within each of these three approaches? Replication failures can be defined in several different ways (e.g., Open Science Collaboration, 2015; Schauer & Hedges, 2021). However, once a methodological decision has been made that identifies the criteria for a replication failure following a corroboration in an original study, that failure can also be taken to provide a logical refutation.[5] Consequently, replication attempts can also represent logical "falsification attempts" (Uygun Tunç et al., 2023, p. 413; see also Sikorski & Andreoletti, 2023, pp. 7–8; Zwaan et al., 2018b, p. 2), and replication failures can represent logical refutations, even if they do not warrant all out theory rejections.

According to my analysis, the NMF approach logically refutes *causal theories*, the Popperian approach logically refutes *noncausal theories*, and the Lakatosian approach logically refutes *noncausal hypotheses*, but not (causal or noncausal) theories. Assuming that logical refutation becomes less scientifically impactful moving from causal theories through noncausal theories to noncausal hypotheses, an unexpectedly large number of logical refutations should be most impactful in the NMF approach, followed by the Popperian approach, with the Lakatosian approach being least affected.

Given that replication failures represent logical refutations, an unexpectedly large number of replication failures should therefore have greatest impact in the NMF approach and moderate impact in the Popperian approaches because these approaches relate to causal and noncausal *theories* respectively. In contrast, an unexpectedly large number of replication failures should be least impactful in the Lakatosian approach because they logically refute only *hypotheses*, not *theories*. Hence, the replication crisis should be less of a "crisis" in Lakatos' philosophy of science than it is in either the NMF or Popperian approaches.

**Table 1** Comparison of the Popperian, Lakatosian, and NMF Approaches

| Approach | Popperian | Lakatosian | NMF |
|---|---|---|---|
| Type of theory | Noncausal | Causal | Causal |
| Example theory | "All swans are white" | "Swanness causes whiteness" | "Swanness causes whiteness" |
| Tentatively accept ceteris paribus clause during testing? | N/A* | No | Yes |
| Logical refutation of theory? | Yes | No | Yes |

*Popper (1974b) explained that "no ceteris paribus clause is necessary" (Footnote 75, pp. 1186–1187)

## 5 Responses to replication failures

Responses to replication failures may also vary depending on the Popperian, NMF, and Lakatosian approach. Here, I consider two responses to direct replication failures: the *hidden moderator* response and the *false positive* response.

One way of explaining a direct replication failure is to assume that the original and direct replication studies are different in some theoretically relevant way. In this case, there can be a correct logical corroboration in the original study (i.e., a true positive result) and a correct logical refutation in the replication study (i.e., a true negative result) because the relevant conditions have changed between the two studies. From this perspective, replication failures are caused by changes in unrecognized factors that were unspecified in the tentatively accepted background knowledge of the original study and that limit the generality of the original result in the direct replication.

This "hidden moderator" response has tended to be dismissed as unrealistic (e.g., Olsson-Collentine et al., 2020, p. 936), lacking in evidence (Kunert, 2016), ad hoc, prone to the hindsight bias, and a "cop-out" (De Ruiter, 2018, p. 17). Moreover, some advocates of direct replications have noted that, "taken to the extreme, this line of reasoning can be used by critics to question the entire enterprise of direct replication" (Zwaan et al., 2018a, p. 46). Certainly, from a Popperian and NMF perspective, it is necessary to tentatively accept the initial conditions, auxiliary hypotheses and, in the case of NMF, ceteris paribus clause as being theoretically equivalent in the original and replication studies in order to logically refute the same theory across the two studies. Zwaan et al. (2018) are correct that the hidden moderator explanation challenges this equivalence and, with it, the key premise of direct replications. From this perspective, it is understandable that some Popperian and NMF researchers may be reticent to adopt the hidden moderator explanation because it precludes a defining feature of their approaches, namely the logical refutation of extant theories. Instead, they may prefer to adopt a false positive explanation because it retains the possibility of such refutations.

The "false positive" explanation assumes that (a) tentatively accepted background knowledge is equally applicable in the original and direct replication studies, (b) the theory was correctly refuted by a direct replication failure, and, consequently, (c) the same theory was incorrectly corroborated in the original study (i.e., a false positive result).[6] The implication of this explanation is that an inadequate theory has slipped through the refutation net due to problems with the theory testing process in the original study, such as nonsevere tests, questionable research practices, and invalid methodology (e.g., Sikorski & Andreoletti, 2023, p. 6). Consequently, the typical response is to try to improve the theory testing process by, for example, tightening up the deductive derivation chain from theory to prediction, using more severe tests, preregistering research plans, using more rigorous methodology, conducting more direct replications, and reducing publication bias.

Unlike the Popperian and NMF approaches, the Lakatosian approach does not attempt to logically refute theories. Consequently, Lakatosians do not face an uncomfortable choice between logically refuting theories and positing hidden moderators. Instead, they are free to get "creative" (Lakatos, 1978, p. 99) by considering potentially relevant differences between the original and replication studies in order to explain replication failures and generating new, falsifiable, "auxiliary hypotheses" (Lakatos, 1978, p. 33) that qualify the "hard core" of their theory (e.g., Lakatos, 1978, p. 179; see also Putnam, 1991, pp. 125–126, 130; for similar reasoning, see Popper, 2002, pp. 56, 62). For example, they might continue to believe that "swanness causes whiteness" but add the auxiliary hypothesis that this causal relation is moderated by location: "swanness causes whiteness, apart from in Australia, where swanness causes blackness" (Karawita et al., 2023; a "boundary condition," Putnam, 1991, pp. 126–127). In this respect, Lakatosians are concerned about the *development* of theories rather than their *logical refutation*.

Based on this approach of iterative theory modification, Lakatos (1978, p. 34) argued that scientists should move away from the appraisal of *single* theories and towards the appraisal of *series* of theories in *research programs*. According to Lakatos, a research program is "a series of theories, $T_1$, $T_2$, $T_3$,…where each subsequent theory results from adding auxiliary clauses to (or from semantical reinterpretations of) the previous theory in order to accommodate some anomaly, each theory having at least as much content as the unrefuted content of its predecessor" (p. 33). Research programs are then assessed in terms of whether they are *progressive* or *degenerative*. In a progressive research program, the new theories accommodate previous anomalies and make new successful predictions. In a degenerating program, however, the new theories only accommodate previous anomalies, and their new predictions remain unsupported (Lakatos, 1978, pp. 34, 179). Hence, from a Lakatosian perspective, theory testing extends across multiple studies, rather than being tied to logical refutations within specific studies. In addition, interpretations of theory testing depend on the knowledge available at specific points in the history of science, and they are open to revision (Lakatos, 1978, pp. 42, 86). Consequently, Lakatosians may also feel free to temporarily ignore replication failures "in the hope that they will turn, in due course, into corroborations of the programme" (Lakatos, 1978, p. 52).

## 6 Alternative perspectives

Like Popper's approach, Lakatos' approach has figured prominently in some discussions of the replication crisis. Here, I consider three alternative perspectives on his approach that have been offered by Zwaan et al. (2018b), Earp and Trafimow (2015), and Uygun-Tunç and Tunç (2023).

Zwaan et al. (2018b) proposed that "replications are an instrument for distinguishing progressive from degenerative research programs" (p. 2; see also Nosek & Errington, 2020, p. 4). However, this proposal is inconsistent with Lakatos' approach (for a similar conclusion, see Fletcher, 2021, p. 4). Lakatosian research programs require studies that test *new* (previously untested) hypotheses of *new*

(previously unknown) effects based on *new* (modified) theories. Hence, they do not imply the *direct* replication attempts that Zwaan et al. advocate. In addition, a Lakatosian research program's negative heuristic forbids the refutation of a theory's hard core (e.g., "swanness causes whiteness"; Lakatos, 1978, p. 48; see also Putnam, 1991, p. 131). Hence, Lakatosian research programs do not even imply *conceptual* replication attempts (i.e., studies that aim to refute the same theoretical hard core under different conditions). Instead, progressive research programs modify and develop theories in order to (a) accommodate previous anomalies and (b) make successful new predictions (e.g., "swanness causes whiteness, apart from in Australia, where swanness causes blackness"). Following Feest (2019, p. 901), the term "exploration" seems more appropriate than "replication" in this context. Furthermore, and contrary to Zwaan et al., it is the results of *innovative new studies*, rather than either direct or conceptual replications, that allow us to distinguish progressive research programs from degenerative ones. As Lakatos (1978) explained:

> "So-called 'refutations' are not the hallmark of empirical failure, as Popper has preached, since all programmes grow in a permanent ocean of anomalies. What really count are dramatic, unexpected, stunning predictions: a few of them are enough to tilt the balance; where theory lags behind the facts, we are dealing with miserable degenerating research programmes" (p. 6).

Earp and Trafimow (2015) also considered the replication crisis in relation to auxiliary hypotheses, ceteris paribus clauses, and Lakatos' (1978) approach. Similar to Zwaan et al. (2018b), they proposed that repeated failures of direct replications by different researchers should gradually decrease confidence in an original study's positive result (Earp & Trafimow, 2015, p. 8). Again, however, following Lakatos' negative heuristic, our confidence in the theoretical hard core that is used to explain a study's positive result should be unaffected by numerous failed replications of that result. Instead, it is our confidence in the progressiveness of a broader research program that should be reduced following the falsification of auxiliary hypotheses that are used to explain replication failures.

Finally, Uygun-Tunç and Tunç's (2023) systematic replications framework (SRF) attempts to reduce underdetermination by distinguishing between potentially relevant and irrelevant auxiliary hypotheses (*AH*s) and then systematically analysing the influence of the potentially relevant *AH*s in a series of close and conceptual replications that relegate the irrelevant *AH*s to a ceteris paribus clause. According to Uygun-Tunç and Tunç, the "SRF is…a tool for assessing if a theory acquires a progressive or degenerative character over time (see Lakatos, 1978)" (p. 5).

A key aspect of the SRF is the relegation of irrelevant *AH*s to the ceteris paribus clause. As Uygun-Tunç and Tunç (2023) explained:

> "Among the plethora of different *AH*s existing in a hypothesis test only a certain subgroup of *AH*s can be expected to meaningfully impact the results. There are infinitely many other *AH*s that presumably do not exert a meaningful

> enough influence on the results due to being completely inconsequential, or only barely consequential so that their influence can be safely ignored to a certain extent, or coinciding with opposing factors that always nullify the potential effect, and so forth. *AH*s that are thought to belong this category are relegated to the ceteris paribus clause (Meehl, 1978). As long as they are deemed to belong to the ceteris paribus clause, they are not explicitly stated, and thus are not tested and (tentatively) accepted as they are" (p. 6).

There are three problems with this approach. First, as Lakatos (1978, p. 18) and Popper (1974b, p. 1187) noted, the role of the ceteris paribus clause is to assume that potentially *relevant* variables are uninfluential during theory testing, not that potentially *irrelevant* variables are uninfluential during testing. In other words, the ceteris paribus clause states that, "although these variables may be influential in the real world, they have no systematic influence on the result of the current test" (see also Meehl, 1990, p. 111). Consequently, it is unnecessary and inconsistent for the SRF to assign *irrelevant* or "inconsequential" variables to a ceteris paribus clause because, by definition, these variables are already assumed to be noninfluential. Only potentially causally relevant (known and unknown) variables need to be assigned to the ceteris paribus clause in order to assume that they are not influential during the theory test.

Second, the SRF operates on the basis of naïve methodological falsificationism because the auxiliary hypotheses that it relegates to the ceteris paribus clause "are not tested and (tentatively) accepted as they are" (Uygun-Tunç & Tunç, 2023, p. 6). As Lakatos (1978) explained, this approach is characteristic of naïve methodological falsificationists, who must decide "whether to relegate also the *ceteris paribus* clause into the pool of 'unproblematic background knowledge'" (p. 26). In contrast, sophisticated methodological falsificationists do not make this decision (Lakatos, 1978, p. 40). For them, the ceteris paribus clause is not tentatively accepted as unproblematic during testing. Instead, *everything* is tested during a theory test, including (a) the theory and (b) auxiliary hypotheses in the ceteris paribus clause (Lakatos, 1978, p. 35).

Third, the SRF is open to the criticism of circularity. As Uygun-Tunç and Tunç (2023) explained, "ultimately all decisions regarding which *AH*s are to be relegated to ceteris paribus clause are theory-laden (See Kuhn, 1996), which might lead one to think that all hypothesis tests are in a way circular" (p. 10). To address this issue, the SRF includes "theory-independent methods for investigating theory misspecification" (p. 10). For example, direct (close) replications represent theory-independent "stability probes":

> "When theory misspecification is due to an erroneous relegation of some crucial *AH*s to the ceteris paribus clause, divergent results in close replications can be an indication of such unspecified *AH*s (e.g., hidden moderators). In this regard, the stability probe is not embedded in the theory under test and constitutes an external success criterion" (Uygun-Tunç & Tunç, 2023, p. 10).

Again, however, this argument demonstrates the SRF's monotheoretical NMF viewpoint because it assumes that there is only a single "theory under test," and that background knowledge, including *AH*s in the ceteris paribus clause, does not

constitute a second *interpretative* theory that is used to determine what counts as "divergent results in close replications" (Uygun-Tunç & Tunç, 2023, p. 10; see Lakatos, 1978, p. 44). From a sophisticated methodological falsificationist perspective, the stability probe *is* embedded in the "theory under test," because it is embedded in the interpretative theory (background knowledge) that is used to interpret the results of the close replications and, as the hidden moderator argument shows, this interpretative theory is just as much under test as the target "explanatory" theory (Lakatos, 1978, p. 44). Consequently, from a Lakatosian perspective, the SRF remains open to the criticism of circularity because its methods for investigating theory misspecification are based on the tentative acceptance of an interpretative theory that may itself be misspecified.

## 7 Conclusion

In conclusion, multiple unexpected replication failures are particularly concerning in the Popperian and NMF approaches because they imply logical refutations of non-causal and causal *theories* respectively. Consequently, scientists' adherence to the Popperian and NMF approaches may be at least partly responsible for the sense of a replication "crisis." In contrast, multiple replication failures are less concerning in the Lakatosian approach because (a) causal theories are not the subject of logical refutations, (b) scientists are used to working in an "ocean of anomalies" (Lakatos, 1978, pp. 6, 53), and (c) replication failures represent opportunities to develop theories rather than to logical refute them (for an illustration, see Sweller, 2023).

## 8 Endnotes

1. In principle, the term *replication failure* may be used to refer to (a) a previously corroborated theory that is falsified in a replication study and (b) a previously falsified theory that is corroborated in a replication study. In line with the standard approach in this area (e.g., Nosek et al., 2022), I reserve the term for the former usage only.
2. In my view, Lakatosian "causal connections" represent what Popper (2002) described as "strictly or purely existential statements (or 'there-is' statements)" (p. 47, italics omitted; e.g., "there is at least one case in which swanness causes whiteness"). Strictly existential statements cannot be falsified by basic statements (Popper, 2002, p. 48). Consequently, Popper treated them as "metaphysical" (p. 48). Confusingly, Popper (1983, p. 288) used the term "causal hypotheses" to refer to "non-probabilistic" hypotheses as opposed to probabilistic hypotheses. However, these non-probabilistic hypotheses did not imply Lakatosian causal connections. It is also worth noting that Popper (2002) replaced the metaphysical principle of causality with a "methodological rule" (p. 39) "always to try to deduce statements from others of higher universality" (p. 107; see also pp. 244–245). For example, one might deduce the statement "all swans are white" from the more universal statement that "all birds are camouflaged" and the "initial

conditions" of swans' historically snowy habitats (Holt, 2022). Again, however, the key point here is that Popperian tests logically refute *noncausal* universal statements rather than *causal* connections. Finally, note that failure to follow Popper's (2002, p. 39) methodological rule "that we are not to abandon the search for universal laws and for a coherent theoretical system" may lead to the charge that one is sliding towards naïve empiricism.

3. Popper used the terms "hypothesis" and "theory" interchangeably. For example, he talked about "the hypothesis (or expectation, or theory, or whatever we may call it)" (Popper, 1974a, p. 346). As a further example, he described the statement "all swans are white" as both a "hypothesis" (e.g., Popper, 1983, p. 343) and a "theory" (e.g., Popper, 1983, p. xx); sometimes on the same page (Popper, 1983, p. 234). Popper also considered the broader concept of a "theoretical system," which contains multiple different hypotheses or theories of varying levels of universality (Popper, 1974b, p. 982; Popper, 2002, pp. 54–55). However, pertinent to the current discussion, (a) theoretical systems may also be logically refuted (Popper, 1974b, p. 982), and (b) any strictly universal statement may be described as a "hypothesis" and a "theory."
4. According to Meehl (1990), "common sense tells us that both the importance and the dangerousness of $C_p$ [a ceteris paribus clause] are much greater in psychology than in chemistry or genetics" (p. 111). In my view, the validity of a ceteris paribus clause should remain in doubt during *any* scientific investigation (see also Trafimow & Fiedler, 2024, p. 8).
5. Popper (2002) argued that researchers should consider a theory "falsified" when they "discover a *reproducible effect* which refutes the theory" (p. 66, italics in original). Hence, in theory, a previously corroborated theory should be logically refuted multiple times to be falsified. However, in response to the question "how often has an effect to be actually reproduced in order to be a 'reproducible effect'," Popper stated "in some cases *not even once*" (Popper, 2002, p. 67, italics in original; see also pp. 23–24). Hence, even a logical refutation in a single replication attempt can falsify a theory when it is based on a severe and independently verifiable test.
6. The "false positive in the original study" explanation is usually based on methodologically rigorous direct replications. Consequently, it tends to play down the possibility of an incorrect refutation in the direct replication (i.e., a false negative result).

**Author contributions** Mark Rubin is the sole author of this manuscript.

**Funding** N/A

**Data availability** N/A

## Declarations

**Informed consent** N/A

**Ethical approval** N/A

**Conflict of interest** N/A



## References

Bird, A. (2021). Understanding the replication crisis as a base rate fallacy. *The British Journal for the Philosophy of Science*, *72*(4), 965–993. https://doi.org/10.1093/bjps/axy051

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces (1st ed.)* Wiley.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E. -J., & Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Chang, A. C., & Li, P. (2022). Is economics research replicable? Sixty published papers from thirteen journals say often not. *Critical Finance Review*, *11*(1), 185–206. https://doi.org/10.1561/104.00000053

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, *66*, 93–99. https://doi.org/10.1016/j.jesp.2015.10.002

De Ruiter, J. P. (2018). The meaning of a claim is its reproducibility. *Behavioral and Brain Sciences*, *41*, e125. https://doi.org/10.1017/S0140525X18000602

Derksen, M. (2019). Putting Popper to work. *Theory & Psychology*, *29*(4), 449–465. https://doi.org/10.1177/0959354319838343

Derksen, M., & Morawski, J. (2022). Kinds of replication: Examining the meanings of conceptual replication and direct replication. *Perspectives on Psychological Science*, *17*(5), 1490–1505. https://doi.org/10.1177/17456916211041116

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, 621, 1–11. https://doi.org/10.3389/fpsyg.2015.00621

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *Elife*, *10*, e71601. https://doi.org/10.7554/eLife.71601

Feest, U. (2019). Why replication is overrated. *Philosophy of Science*, *86*(5), 895–905. https://doi.org/10.1086/705451

Feyerabend, P. (1975). Imre Lakatos. *The British Journal for the Philosophy of Science*, *26*(1), 1–18. https://doi.org/10.1093/bjps/26.1.1

Feynman, R. P. (1955). The value of science. *Engineering and Science*, *19*(3), 13–15. https://calteches.library.caltech.edu/1575/1/Science.pdf

Firestein, S. (2012). *Ignorance: How it drives science*. Oxford University Press.

Fletcher, S. C. (2021). The role of replication in psychological science. *European Journal for Philosophy of Science*, *11,* 23, 1–19. https://doi.org/10.1007/s13194-020-00329-2

Flis, I. (2019). Psychologists psychologizing scientific psychology: An epistemological reading of the replication crisis. *Theory & Psychology*, *29*(2), 158–181. https://doi.org/10.1177/0959354319835322

Forscher, B. K. (1963). Chaos in the brickyard. *Science*, *142*(3590), 339–339. https://doi.org/10.1126/science.142.3590.339.a

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802. https://doi.org/10.1177/1745691620970585

Hager, W. (2000). About some misconceptions and the discontent with statistical tests in psychology. *Methods of Psychological Research Online*, *5*(1), 1–31. https://psycnet.apa.org/record/2001-03459-001

Haig, B. D. (2009). Inference to the best explanation: A neglected approach to theory appraisal in psychology. *The American Journal of Psychology*, *122*(2), 219–234. https://doi.org/10.2307/27784393

Holt, D. W. (2022). Why are snowy owls white and why have they evolved distinct sexual color dimorphism? A review of questions and hypotheses. *Journal of Raptor Research*, *56*(4), 440–454. https://doi.org/10.3356/JRR-21-56

Johansson, I. (1980). Ceteris paribus clauses, closure clauses and falsifiability. *Zeitschrift für Allgemeine Wissenschaftstheorie*, *11*, 16–22. https://doi.org/10.1007/BF01801276

Karawita, A. C., Cheng, Y., Chew, K. Y., Challagulla, A., Kraus, R., Mueller, R. C., Tong, M. Z. W., Hulme, K. D., Bielefeldt-Ohmann, H., Steele, L. E., Wu, M., Sng, J., Noye, E., Bruxner, T. J., Au, G. G., Lowther, S., Blommaert, J., Suh, A., McCauley, A. J., Kaur, P., Dudchenko, O., Aiden, E., Fedrigo, O., Formenti, G., Mountcastle, J., Chow, W., Martin, F. J., Ogeh, D. N., Thiaud-Nissen, F., Howe, K., Tracey, A., Smith, J., Kuo, R. I., Renfree, M. B., Kimura, T., Sakoda, Y., McDougall, M., Spencer, H. G., Pyne, M., Tolf, C., Waldenström, J., Jarvis, E. D., Baker, M. L., Burt, D. W., & Short, K. R. (2023). The swan genome and transcriptome, it is not all black and white. *Genome Biology*, *24*, 13, 1–24. https://doi.org/10.1186/s13059-022-02838-0

Kunert, R. (2016). Internal conceptual replications do not increase independent replication success. *Psychonomic Bulletin & Review*, *23*(5), 1631–1638. https://doi.org/10.3758/s13423-016-1030-9

Lakatos, I. (1968). Criticism and the methodology of scientific research programmes. *Proceedings of the Aristotelian Society*, *69*, 149–186. https://www.jstor.org/stable/4544774

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos, & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91–196). Cambridge University Press.

Lakatos, I. (1978). *The methodology of scientific research programmes (Philosophical Papers, Volume I)*. Cambridge University Press.

Maziarz, M. (2024). Conflicting results and statistical malleability: Embracing pluralism of empirical results. *Perspectives on Science*, *32*(6), 701–728. https://doi.org/10.1162/posc_a_00627

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1

Merton, R. K. (1987). Three fragments from a sociologist's notebooks: Establishing the phenomenon, specified ignorance, and strategic research materials. *Annual Review of Sociology*, *13*(1), 1–29. https://doi.org/10.1146/annurev.so.13.080187.000245

Monnerjahn, P. (2019). A review of "Statistical Inference as Severe Testing". *The Open Society: Enlightening Ideas, Critical Discussion.* http://www.theopensociety.net/a-review-of-statistical-inference-as-severe-testing/

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E. J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 0021. https://doi.org/10.1038/s41562-016-0021

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLoS biology*, *18*(3), e3000691. https://doi.org/10.1371/journal.pbio.3000691

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*, 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

O'Donohue, W. (2021). Are psychologists appraising research properly? Some Popperian notes regarding replication failures in psychology. *Journal of Theoretical and Philosophical Psychology*, *41*(4), 233–247. https://doi.org/10.1037/teo0000179

Olsson-Collentine, A., Wicherts, J. M., & van Assen, M. A. L. M. (2020). Heterogeneity in direct replications in psychology and its association with effect size. *Psychological Bulletin*, *146*(10), 922–940. https://doi.org/10.1037/bul0000294

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Pearce, N. (1990). White swans, black ravens, and lame ducks: Necessary and sufficient causes in epidemiology. *Epidemiology (Cambridge, Mass.)*, *1*, 47–50. http://www.jstor.org/stable/20065623

Popper, K. R. (1974a). *Objective knowledge: An evolutionary approach*. Oxford University Press.

Popper, K. R. (1974b). Reply to my critics. In P. A. Schilpp (Ed.), *The philosophy of Karl Popper (Book II)* (pp. 960–1197). Open Court.

Popper, K. R. (1983). *Realism and the aim of science: From the postscript to the logic of scientific discovery*. Routledge.

Popper, K. R. (2002). *The logic of scientific discovery*. Routledge.

Putnam, H. (1991). The 'corroboration' of theories. In R. Boyd, P. Gasper, & J. D. Trout (Eds.), *The philosophy of science* (pp. 121–137). MIT Press.

Reutlinger, A., Schurz, G., Hüttemann, A., & Jaag, S. (2021). Ceteris paribus laws. *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition). https://plato.stanford.edu/archives/fall2021/entries/ceteris-paribus

Rubin, M. (2020). "Repeated sampling from the same population?" A critique of Neyman and Pearson's responses to Fisher. *European Journal for Philosophy of Science*, *10*, 42, 1–15. https://doi.org/10.1007/s13194-020-00309-6

Rubin, M. (2023b). Questionable metascience practices. *Journal of Trial and Error*, *4*(1), 5–20. https://doi.org/10.36850/mr4

Rubin, M. (2024). Type I error rates are not usually inflated. *Journal of Trial and Error, 4*(2), 46–71. https://doi.org/10.36850/4d35-44bd

Rubin, M. (2023a, June 7). The preregistration prescriptiveness trade-off and unknown unknowns in science: Comments on Van Drimmelen (2023). *MetaArXiv*. https://doi.org/10.31222/osf.io/3t7pc

Schauer, J. M., & Hedges, L. V. (2021). Reconsidering statistical methods for assessing replication. *Psychological Methods*, *26*(1), 127–139. https://doi.org/10.1037/met0000302

Sikorski, M., & Andreoletti, M. (2023). Epistemic functions of replicability in experimental sciences: Defending the orthodox view. *Foundations of Science*, 29, 1071–1088. https://doi.org/10.1007/s10699-023-09901-4

Strong, S. R. (1991). Theory-driven science and naïve empiricism in counseling psychology. *Journal of Counseling Psychology*, *38*(2), 204–210. https://doi.org/10.1037/0022-0167.38.2.204

Sweller, J. (2023). The development of cognitive load theory: Replication crises and incorporation of other theories can lead to theory expansion. *Educational Psychology Review*, *35*, 95, 1–20. https://doi.org/10.1007/s10648-023-09817-2

Trafimow, D., & Fiedler, D. (2024). An exploration of physics envy with implications for desiderata of psychology theories. *American Psychologist*. https://doi.org/10.1037/amp0001416

Uygun Tunç, D., Tunç, M. N., & Lakens, D. (2023). The epistemic and pragmatic function of dichotomous claims based on statistical hypothesis tests. *Theory & Psychology*, *33*(3), 403–423. https://doi.org/10.1177/09593543231160112

Uygun-Tunç, D., & Tunç, M. N. (2023). A falsificationist treatment of auxiliary hypotheses in social and behavioral sciences: Systematic replications framework. *Meta-Psychology, 7*. https://doi.org/10.15626/MP.2021.2756

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018a). Improving social and behavioral science by making replication mainstream: A response to commentaries Responses. *Behavioral and Brain Sciences*, *41*, e157. https://doi.org/10.1017/S0140525X18000961

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018b). Making replication mainstream. *Behavioral and Brain Sciences*, *41*, e120. https://doi.org/10.1017/S0140525X17001972