

What a Failed Replication Proves

A finding fails to reproduce in Utrecht. Both laboratories behaved properly. The disagreement that follows is not about the data.

12 min read

Before you read

- **replication** — an independent repetition of a study intended to test its finding
- **effect size** — the magnitude of a difference, as distinct from its statistical significance
- **statistical power** — the probability that a study would detect an effect if one exists
- **auxiliary assumption** — a background premise a test depends on but does not itself test
- **moderator** — a variable that changes the size or direction of an effect
- **ad hoc** — introduced to rescue a claim from a particular counter-example
- **falsifiable** — capable in principle of being shown false by evidence
- **underdetermination** — the situation in which evidence is compatible with rival explanations



Testing cubicles in a university behavioural laboratory, Utrecht

¹ In 2014 a laboratory in Utrecht attempted to reproduce a finding that had been cited several hundred times: that participants primed with words associated with age subsequently walked more slowly down a corridor. The original study reported a clear effect in a sample of thirty. The replication used three hundred participants, pre-registered its analysis, and found nothing: an effect size indistinguishable from zero, with confidence intervals narrow enough to exclude the original estimate.

² The sequence that followed has been repeated, with variations, across a decade of psychology, economics, medicine and ecology, and it is worth examining structurally rather than as a dispute between

individuals. Neither laboratory did anything improper. That is what makes the case instructive.

What the replication can and cannot show

³ A failed replication is a negative result, and negative results are asymmetrically weak. A study that finds an effect establishes that something produced a difference under those conditions. A study that finds no effect establishes only that it did not detect one — which is consistent with the effect being absent, with it being smaller than the study was powered to detect, and with it being present but dependent on a condition the replication did not reproduce.

⁴ The Utrecht study is stronger than most negative results precisely because of its precision. Three hundred participants and a pre-registered analysis produce intervals tight enough that the original effect, had it been real and of the reported magnitude, would almost certainly have appeared. This moves the question. The live possibility is no longer that the effect was missed; it is that the effect is substantially smaller than reported, or that it depends on something unstated.

The original author's reply

⁵ The original author responded that the replication differed in several respects: the corridor was shorter, the experimenters were not blind to condition in the same way, and the participant pool was Dutch rather than American. Each of these is a proposed moderator — a variable on which the effect is said to depend. The reply is not unreasonable. Context sensitivity is a real feature of behavioural research, and an effect that holds only under certain conditions is not thereby a non-effect.

⁶ But the reply has a structural cost that is easy to miss. The moderators were not specified in advance. They were identified after the failure, selected from the indefinitely large set of ways in which any two studies differ. A defence constructed this way will always be available, because two laboratories always differ in some respects, and it therefore does no work in distinguishing a real context-dependent effect from no effect at all. This is what philosophers of science mean by an ad hoc rescue: not a false claim, but one whose availability is guaranteed in advance of the evidence.

Why neither side can be purely at fault

⁷ The deeper problem is that no hypothesis is tested alone. Testing "age primes slow walking" requires auxiliary assumptions: that the prime was perceived, that walking speed was measured in a way that reflects the construct, that the sample is relevantly similar, that the laboratory context does not suppress the effect. A negative result shows that something in the conjunction is false. It does not identify which element. This is underdetermination, and it is not a flaw in psychology — it is a general feature of empirical testing, visible in physics whenever an anomalous measurement is attributed to the apparatus rather than to the theory.

⁸ What distinguishes a legitimate appeal to auxiliaries from an illegitimate one is not the content of the appeal but what follows it. The physicist who blames the apparatus is expected to specify the fault and show that correcting it restores the prediction. The equivalent move here would be to state the moderator precisely, predict in advance where the effect should and should not appear, and run that study. Where that happens, the rescue has become a hypothesis. Where it does not, the claim has been removed from the reach of evidence while retaining the appearance of an empirical finding.

The incentive layer

⁹ None of this is primarily about individual honesty, and treating it that way has made the literature worse rather than better. Small samples were not a secret; they were the professional norm, rewarded by journals that published surprising positive findings and declined negative ones. A researcher following every rule of their field in 2008 would produce exactly the body of work now failing to replicate. The reforms that have mattered — pre-registration, published protocols, registered reports accepted before results are known, routine power analysis — are structural rather than moral. They change what a careful researcher is able to do, which is more reliable than changing what a careless one intends.

¹⁰ The uncomfortable conclusion is that the Utrecht result does not settle the question it was designed to settle, and could not have. What it does is shift the burden. Before it, the claim was supported by a published finding and cost nothing to believe. After it, continued belief requires a specific, testable account of why the effect appears in one corridor and not another. That is not proof of absence. It is the ordinary mechanism by which a literature either sharpens a claim or loses it, and the mechanism works only if someone is willing to pay the cost of the negative result.

Part A **Select the best answer**

Circle the letter next to the answer best supported by the text.

1. The author describes negative results as "asymmetrically weak" because they
 - A are rejected more often by journals than positive results
 - B cannot distinguish absence of an effect from failure to detect one
 - C require larger samples than positive results
 - D depend on auxiliary assumptions that positive results avoid
2. The Utrecht study is said to be stronger than most negative results because
 - A it was conducted by an independent laboratory
 - B it used the same corridor length as the original
 - C its precision excludes the originally reported effect size
 - D it was published alongside the original author's reply
3. The author's objection to the proposed moderators is that they were
 - A implausible as influences on walking speed
 - B already tested and rejected in the original study
 - C selected after the failure from an unlimited set of differences
 - D too numerous to test within a single experiment
4. The comparison with physics in paragraph 7 is introduced in order to
 - A establish that underdetermination is general, not a defect of psychology
 - B argue that physical measurements are more reliable
 - C show that auxiliary assumptions are rarely false
 - D criticise psychologists for borrowing physical methods

Part B **Analyse and explain**

Answer in continuous prose. Refer to the text by paragraph number where it strengthens your point.

5. Explain why the author insists the case be examined "structurally rather than as a dispute between individuals" (paragraph 2). What is gained?

6. Set out the conditions under which, on the author's account, an appeal to a moderator stops being ad hoc.
- _____
- _____
- _____
- _____
- _____
7. Explain underdetermination as the passage uses it, and state what it implies about what the Utrecht result licenses anyone to conclude.
- _____
- _____
- _____
- _____
- _____
8. Paragraph 9 argues the reforms that mattered were "structural rather than moral". Explain the distinction and why the author prefers the former.
- _____
- _____
- _____
- _____
- _____
9. Analyse the final sentence. What does "pay the cost" refer to, and how does it connect to the argument of paragraph 9?
- _____
- _____
- _____
- _____

Part C **Argue a position**

There is no single correct answer. Take a position, and make the reasoning that supports it visible.

10. The author claims the Utrecht result "shifts the burden" rather than settling the question. Is burden-shifting a satisfactory outcome for an expensive study? Argue a position.

11. Design the study paragraph 8 calls for: state a specific moderator, and the prediction that would distinguish a real context-dependent effect from no effect. Explain why your design has that property.

12. If a researcher following every norm of their field produces unreliable work, where does responsibility lie? Make the case for at least two candidates before settling on one.

13. Pre-registration constrains researchers in advance. Construct the strongest objection to it — from exploratory research, say — and then evaluate that objection.

14. The passage assumes that a claim removed from the reach of evidence has lost something important. Is that always so? Consider whether any valuable claims are legitimately unfalsifiable.

Answer key

For parents and teachers

Part A

1. B | 2. C | 3. C | 4. A

Part B

5. Treating it as a personal dispute makes the lesson unrepeatable and invites blame in place of diagnosis; treating it structurally shows that the same sequence arises wherever the same incentives and testing logic operate, which is what makes reform possible. Strong answers link this to paragraph 9's claim that a rule-following researcher produces the same result.

6. The moderator must be stated precisely rather than gestured at, specified in advance of the evidence it is meant to explain, used to generate a prediction about where the effect will and will not appear, and then tested. At that point it has become a hypothesis rather than a rescue. Credit answers that capture advance specification plus testability.

7. A hypothesis is never tested alone but always in conjunction with auxiliary assumptions, so a negative result shows only that something in the conjunction is false, not which element. The Utrecht result therefore licenses no conclusion about the priming hypothesis specifically — only about the whole package, unless further evidence isolates the failing component.

8. Moral reform asks individuals to intend better; structural reform changes what the system makes publishable and therefore what careful work looks like. The author prefers the structural because the failure was produced by norm-following rather than norm-breaking, so appeals to intention address the wrong cause.

9. The cost is professional: negative results are expensive to run, hard to publish, and attract conflict rather than citation. It connects to paragraph 9 because the incentive structure that produced the unreliable literature also discourages the work that corrects it — so the mechanism described is not self-executing.

Part C

These answers vary. Mark the reasoning, not the conclusion.

10. Either position is markable. Strong answers weigh the value of narrowing a live hypothesis against the opportunity cost, and notice that the author's framing makes burden-shifting the normal, not the disappointing, outcome of science.

11. Mark the design property, not the moderator chosen: the prediction must be able to fail. A design that predicts an effect somewhere but excludes it nowhere earns little credit.

12. Candidates include journals, funders, supervisors, the individual, and the discipline's methodological training. Reward genuine engagement with at least two before a reasoned choice.

13. The strongest objection is that exploratory work legitimately discovers what it was not looking for, and pre-registration can push genuine discovery into disguised confirmation. Look for the standard reply: label it exploratory and confirm it separately.

14. Strong answers distinguish claims that are unfalsifiable and therefore empty from claims that are not empirical at all — mathematical, definitional or ethical — and notice that the author's complaint is about disguise, not about unfalsifiability as such.

EyeQBoost builds thinking skills in young learners through cognitive training, reading and gamified online classes.
More free reading sets at eyeqboost.com/resources.