

When the Experiment Will Not Repeat

A famous finding, a careful re-run, and a question a discipline could not avoid asking itself.

7 min read

Before you read

- **replication** — re-running a study to see whether the result appears again
- **null result** — a finding that shows no effect
- **pre-registration** — publishing your hypothesis and analysis plan before collecting data
- **significance threshold** — the agreed probability below which a result is treated as real
- **incentive structure** — the rewards a system gives, which shape the behaviour inside it
- **self-correction** — a field identifying and repairing its own errors



A row of identical laboratory cubicles in a Dutch university, one occupied, the rest empty

¹ In the autumn of 2011, three junior researchers at a Dutch university did something their training had not prepared them for. They went to the dean and reported that a senior colleague's data looked wrong. Not wrong in the ordinary way — too tidy. The numbers fell into patterns that real human beings, with their inconsistencies and their bad days, do not produce.

² The investigation that followed found that dozens of published papers rested on data that had never been collected. The case was extreme, and the discipline was quick to say so. Fraud is rare. Most researchers are honest. Psychology, it was argued, had caught the problem and corrected itself.

³ That reassurance did not survive the decade. The deeper problem turned out not to be fraud at all, but something far more ordinary and considerably harder to legislate against.

⁴ From 2011 onwards, groups of psychologists set out to repeat well-known experiments exactly as published. The logic was simple: a real effect should appear again when the conditions are reproduced. In 2015 the Open Science Collaboration reported the outcome of a hundred such attempts. Fewer than half produced the original effect, and where an effect did appear it was typically about half the size.

⁵ The findings that failed were not obscure. They included results that had reached textbooks, conference stages and corporate training rooms: that holding a confident posture alters hormone levels, that thinking about old age makes people walk more slowly, that willpower is a finite resource depleted by use.

⁶ Nobody had fabricated these. They came out of laboratories doing exactly what the discipline had taught them to do. That is the uncomfortable part.

⁷ Consider the ordinary life of an experiment. A researcher collects data from forty participants and finds nothing. Rather than abandon months of work she adds twenty more — a reasonable response to a small sample. Now the result is significant. She measured four outcomes, and reports the one that worked, since the others seem uninformative. She divides her participants by sex, because the pattern is clearer among the women. Each decision is defensible in isolation. Taken together, they manufacture a finding out of noise.

⁸ Statisticians call this the garden of forking paths. The researcher never lies about what she did; she simply chooses, at each small junction, the route the data recommends. Because those choices are made after the data is in, the conventional significance threshold — the familiar probability of one in twenty — no longer means what it claims to mean. A test designed to be wrong five times in a hundred can, given enough flexibility, be wrong a third of the time.

⁹ The incentive structure did the rest. Journals published surprising positive results and rejected null ones. Hiring committees counted publications. A failed replication was not a contribution but an embarrassment, and frequently an accusation. Nobody was rewarded for checking.

¹⁰ What has changed since is largely procedural, and more consequential than that sounds. Researchers now pre-register their hypotheses and analysis plans before collecting data, which closes the forking paths by fixing the route in advance. Some journals accept papers on the strength of the method alone, before any results exist — a format under which publication no longer depends on the answer. Data and code are posted openly. Sample sizes have grown.

¹¹ The reforms carry costs, and they are not trivial. Pre-registration constrains the exploratory work from which genuinely new hypotheses often come. Larger samples are expensive, and that expense falls hardest on researchers at poorer institutions, which risks concentrating publishable science in wealthy universities. And a field that advertises its own unreliability invites a reading it does not intend: that if psychology cannot replicate, no social science can be trusted at all.

¹² That last conclusion is the one worth resisting. A discipline that audits itself in public and publishes the result is behaving better, not worse, than one that declines to look. The replication crisis is evidence of a

correction already under way. Whether it is a sufficient correction is a different question, and the honest answer is that we will not know for another decade — roughly the time it takes for a cohort trained under the new rules to fill the journals.

Part A Reading closely

Circle the letter next to the best answer.

1. What did the 2015 Open Science Collaboration report find?
 - A All one hundred experiments were shown to be fraudulent
 - B Fewer than half of the hundred attempts reproduced the original effect
 - C The original effects were confirmed but found to be larger than reported
 - D Only experiments conducted in the Netherlands failed to replicate
2. In paragraph 8, the phrase *the garden of forking paths* refers to
 - A the deliberate fabrication of data by dishonest researchers
 - B the difficulty of recruiting enough participants for a study
 - C the series of defensible analytic choices a researcher makes after seeing the data
 - D the practice of publishing the same finding in several journals
3. According to paragraph 10, what distinguishes the publication format described there from a conventional paper?
 - A It is accepted on the strength of its method, before the results exist
 - B It can only be submitted by researchers at wealthy institutions
 - C It is published without the data or code being made available
 - D It reports null results only
4. Which of the following does the writer identify as a cost of the reforms?
 - A Journals have begun to reject replication attempts outright
 - B Significance thresholds have become impossible to calculate
 - C Fraud has become harder to detect than it was in 2011
 - D The expense of larger samples falls hardest on poorer institutions

Part B Analysis

Answer in two or three sentences each. Refer to the passage by paragraph number where it helps your answer.

5. Paragraph 3 claims the deeper problem was ‘far more ordinary’ than fraud. Explain what the writer means, and why ordinariness makes the problem harder to solve.
6. The researcher described in paragraph 7 never lies. Why, then, does the writer treat her finding as unreliable?
7. Explain the role the writer assigns to journals and hiring committees in paragraph 9. What is being blamed here, and what is not?
8. Paragraph 11 concedes three costs of reform before paragraph 12 defends it. What does making those concessions do for the strength of the argument?

9. Comment on the effect of the final clause of the passage, with its estimate of a decade. How does it shape the tone on which the piece ends?

Part C Think deeper

There is no single right answer here. Take a position and support it with reasoning and, where relevant, evidence from the passage.

10. The writer treats public self-auditing as a sign of health. Construct the strongest opposing case: that publicising failure damages a field more than the errors themselves do.

11. Pre-registration closes the forking paths but constrains exploratory work. Which risk should a developing field be more willing to accept, and why?

12. Apply the forking-paths problem to a field outside psychology — medicine, economics, education, sport science. Where would the same pressures produce the same result?

13. The passage argues that a system rewarded nobody for checking. Design one change to that incentive structure, and explain both what it would fix and what it would cost.

14. 'A result that has not been replicated should not be taught.' Argue for or against this as a rule for textbooks and classrooms.

Answer key

For parents and teachers

Part A

1. **B** — fewer than half reproduced the original effect (para 4) | 2. **C** — defensible choices made after seeing the data (para 8)
3. **A** — accepted on method, before results exist (para 10) | 4. **D** — cost of larger samples falls on poorer institutions (para 11)

Part B

5. Fraud is deliberate, rare and removable: you dismiss the individual and the problem leaves with them. The deeper problem is routine analytic flexibility practised by honest researchers following accepted training, so there is no wrongdoer to remove and no single rule to break. Credit any answer that grasps the contrast between a rare deliberate act and a widespread ordinary practice.
6. Because the choices were made after she saw the data. Adding participants, selecting the outcome that worked and splitting the sample are each defensible alone, but together they inflate the chance of a false positive well beyond the stated one in twenty. Honesty about method is not the same as validity of result — strong answers make that distinction.
7. The incentive structure is blamed, not the individuals operating inside it. Journals rewarded surprising positive findings, hiring committees counted publications, and replication attempts were treated as accusations, so nobody was rewarded for checking. Accept answers noting that the writer deliberately locates fault in the system rather than in character.
8. Conceding real costs pre-empts the obvious objections and establishes the writer as even-handed rather than partisan, so the defence in paragraph 12 reads as a judgement reached rather than a position held. Credit answers that name the rhetorical function — concession strengthening credibility — rather than merely listing the three costs.
9. It withholds the resolution the reader expects. The piece ends on measured uncertainty rather than triumph, and the mechanism given — waiting for a trained cohort to fill the journals — is deliberately unglamorous. Accept: sober, honest, anticlimactic, self-consistent with the argument for patience and verification.

Part C

These answers vary. Mark the reasoning, not the position taken.

10. A strong answer builds the opposing case properly rather than dismissing it: public failure is seized on by bad-faith critics, erodes funding, and discourages students from entering the field, so reform conducted quietly might preserve more of the enterprise. A weak answer simply restates the writer's position and calls the alternative wrong.
11. Look for an actual trade-off being weighed, with a criterion named — maturity of the field, cost of a false positive, availability of replication funding. Answers that only assert one side without acknowledging what is lost have not done the work.
12. The transfer is the point. Strong answers identify the specific mechanism in the chosen field — trial endpoints switched in medicine, specification choices in economics, outcome measures in education — rather than saying only that the field also publishes research.
13. Credit a concrete proposal with a named cost: funding replication directly, weighting hiring on fewer and better papers, journals committing to publish null results. Any proposal offered without its cost is an incomplete answer at this level.
14. Either side can be argued well. The strongest answers distinguish between teaching a finding as established and teaching it as contested, which dissolves much of the apparent conflict. Prompt with: what would be lost from the curriculum if the rule were applied strictly?

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