INQUIRY · LOGIC AND REASONING

Evaluating the Quality of an Inquiry, Not Just Its Answer

A correct answer can come from a poor inquiry, and a sound inquiry can produce a wrong one. This article explains why the quality of an inquiry should be judged separately from its result, the criteria researchers use to do so, and how organisations can apply them.

ModQu research article · 2,143 words · 24 references (APA 7th edition)

When a piece of research turns out to be right, it is natural to conclude that it was good research. When it turns out to be wrong, it is natural to conclude that it was bad. Both conclusions can be mistaken. A forecast may come true by luck; a carefully designed study may be overturned by an event no one could have foreseen. If organisations judge their inquiries only by whether the answers turned out to be right, they will reward lucky guesses, punish sound work and learn the wrong lessons. This article explains why the quality of an inquiry should be evaluated separately from its answer, what criteria researchers in different fields use, and how those criteria can be applied in practice.

Why the answer is not enough

Three problems make the answer an unreliable guide to the quality of the inquiry that produced it.

Luck. In any uncertain domain, a sound process will sometimes produce a wrong answer and a weak process will sometimes produce a right one. Annie Duke (2018), drawing on her experience as a professional poker player, called the habit of judging decisions by their results "resulting". A poker player who wins a hand after a bad decision has not played well; one who loses after a good decision has not played badly. Over many hands, good decisions win, but any single result says little.

Outcome bias. Jonathan Baron and John Hershey (1988) showed that people rate the same decision as better when they are told it turned out well, even when they know the outcome could not have been foreseen. The bias affects evaluations of doctors, managers and researchers alike. Baruch Fischhoff (1975) found a related effect: once people know the outcome, they believe it was more predictable than it was. Together, these biases make an inquiry that got the right answer look wise and one that got the wrong answer look careless.

Delay. The answer to many inquiries cannot be checked for months or years, and some can never be checked directly. Decisions have to be made long before the verdict is in. If quality can only be judged by outcomes, it cannot be judged in time to matter.

Figure 1 shows the four combinations of process and outcome. The cell that matters most is the one in which a weak inquiry produces a good result, because it teaches the organisation to repeat a process that will eventually fail.

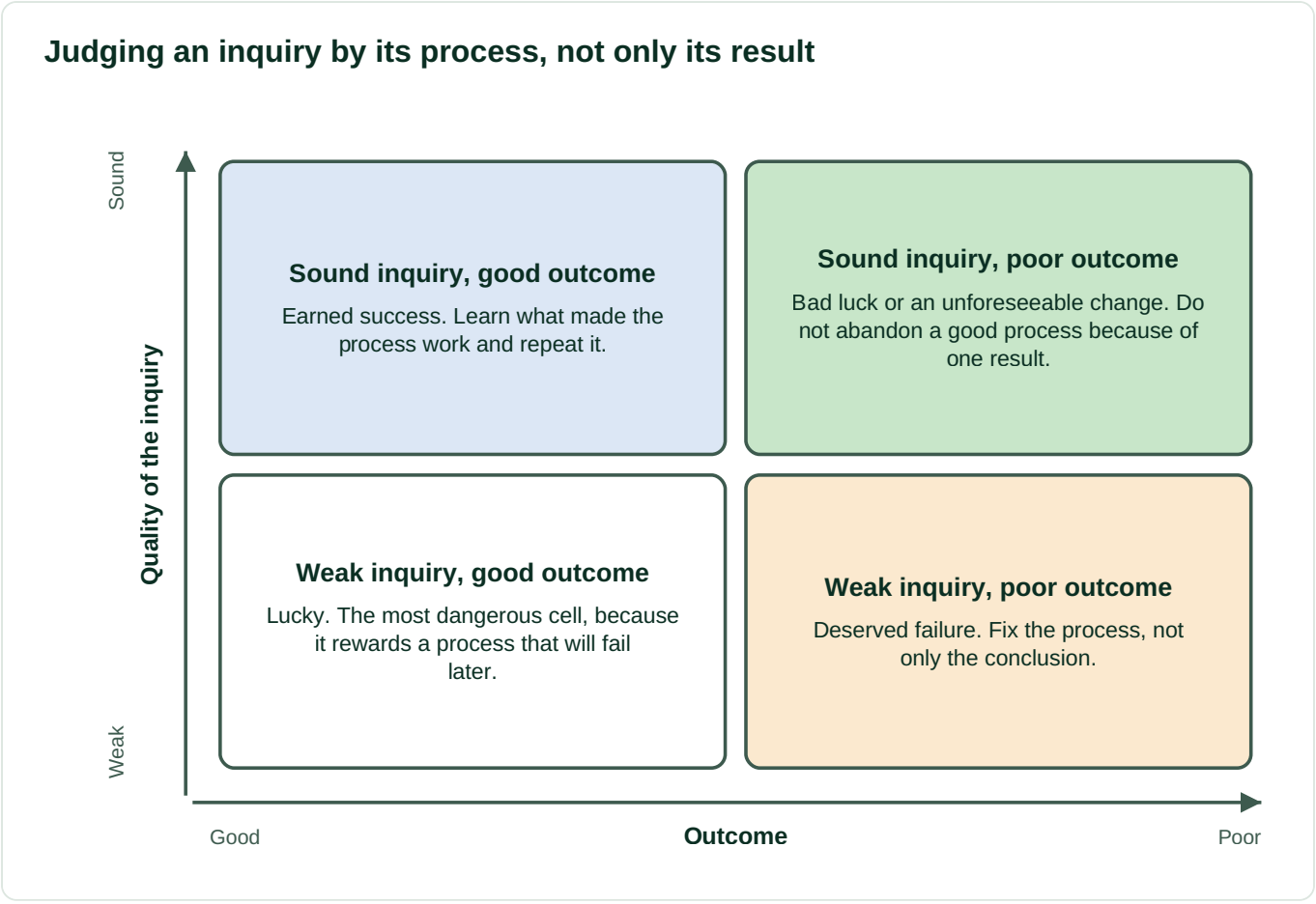


Figure 1. Process and outcome. Only the quality of the inquiry is under the inquirer's control. Diagram by ModQu.

Philosophical foundations

The idea that the quality of a belief depends on how it was formed has deep roots. John Dewey (1938) argued that the outcome of inquiry should be called "warranted assertibility" rather than truth, to emphasise that a conclusion is justified by the process of inquiry that supports it. A conclusion reached by careful inquiry is warranted even if later evidence overturns it, and a true conclusion reached by guesswork is not.

The philosopher Alvin Goldman (1979) developed a related account of justification known as *reliabilism*. A belief is justified, on this view, if it was produced by a process that tends to produce true beliefs. A lucky guess that happens to be true is not justified, because guessing is not a reliable process. Applied to inquiry, this suggests evaluating research methods by their track record across many cases, not by the result of any single study.

Karl Popper (1959) contributed a further criterion: a good inquiry exposes its conclusions to the possibility of being wrong. A theory or finding that could not have been contradicted by any possible evidence has not been tested at all. Deborah Mayo (2018) developed this idea into the principle of *severe testing*: a claim is well supported only if it has passed a test that would probably have caught it if it were false. An inquiry that "confirms" a hypothesis using a method that would have confirmed it whatever the truth gives no real support.

How research fields judge quality

Different research traditions have developed explicit criteria for judging the quality of inquiry independently of its conclusions.

Validity in quantitative research. Shadish et al. (2002) set out four types of validity for studies of cause and effect. *Statistical conclusion validity* concerns whether the statistical analysis supports the conclusion. *Internal validity* concerns whether the observed relationship is causal. *Construct validity* concerns whether the study measured and manipulated what it claims to have. *External validity* concerns whether the results generalise to other people, settings and times. Samuel Messick (1995) argued that the validity of a measurement depends on the evidence and reasoning supporting its interpretation and use, not just on the instrument itself.

Trustworthiness in qualitative research. Yvonna Lincoln and Egon Guba (1985) proposed criteria suited to qualitative inquiry: *credibility* (are the findings believable to those studied?), *transferability* (can readers judge where they apply?), *dependability* (is the process documented and consistent?) and *confirmability* (are the findings grounded in the data rather than the researcher's preferences?). Sarah Tracy (2010) added a broader set of eight criteria, including a worthy topic, rich rigour, sincerity, credibility and meaningful coherence between the question, methods and conclusions.

Evidence grading in medicine. Medicine has developed formal systems for judging the quality of evidence before deciding what it means. The GRADE system rates the certainty of a body of evidence by considering risk of bias, inconsistency between studies, indirectness, imprecision and publication bias (Guyatt et al., 2008). The Cochrane risk-of-bias tool asks specific questions about how a trial was conducted, such as how participants were allocated to groups and whether outcome data are missing (Higgins et al., 2011). These systems judge the quality of a study by its methods, deliberately independent of whether its results are welcome.

Research programmes. Imre Lakatos (1970) proposed judging whole programmes of research over time. A *progressive* programme predicts new facts that are later confirmed. A *degenerating* programme only explains facts after the event, adding assumptions to accommodate each new failure. The distinction applies to an organisation's ongoing research as well as to scientific theories.

Lessons from the replication crisis

Recent experience in science shows what happens when quality is judged by results. For decades, journals favoured studies with striking, statistically significant findings. John Ioannidis (2005) argued that this, combined with small samples and flexible analysis, meant that many published findings were likely to be false. Simmons et al. (2011) showed that undisclosed choices in data collection and analysis could make almost any result appear significant. Andrew Gelman and Eric Loken (2014) described how even researchers acting in good faith can be misled by the "garden of forking paths":
the many analysis decisions that could reasonably have been made differently, each affecting the result. When the Open Science Collaboration (2015) attempted to replicate 100 psychology studies, far fewer produced significant results the second time.

The response has been to shift attention from results to process. Pre-registration, in which researchers record their hypotheses and analysis plans before seeing the data, makes it possible to distinguish tests planned in advance from findings discovered along the way (Nosek et al., 2018). Some journals now review and accept studies on the basis of their question and method, before the results are known, so that publication depends on the quality of the inquiry rather than the excitement of its answer.

Judging forecasters and analysts

Where judgements are made repeatedly, process quality can be measured through track records. Philip Tetlock and Dan Gardner (2015) describe forecasting tournaments in which thousands of participants made probability forecasts about world events, which were scored once the outcomes were known. Scoring many forecasts, rather than one, separates skill from luck. Mellers et al. (2014) found that the best forecasters were distinguished not by special knowledge but by how they worked: they updated their beliefs frequently in small steps, considered multiple perspectives and were trained in basic probabilistic reasoning. Working in teams also improved accuracy.

The lesson for organisations is that an analyst's quality cannot be judged from one high-profile call. It can be judged from many forecasts, recorded at the time and scored afterwards, and from the methods the analyst uses.

Process matters for decisions too

Evidence from business supports the same conclusion. Lovallo and Sibony (2010) surveyed managers about more than 1,000 major investment decisions and found that the quality of the decision process, such as whether uncertainties were discussed openly and dissenting views heard, explained much more of the variation in outcomes than the quality of the quantitative analysis. Kahneman et al. (2011) proposed that leaders reviewing a recommendation should examine how the recommendation was produced, using a checklist of questions about possible biases, such as whether the team considered credible alternatives and whether it was overly attached to its own proposal.

Five dimensions of inquiry quality

Drawing these strands together, the quality of an inquiry can be judged along five dimensions, summarised in Figure 2.

Five dimensions of inquiry quality

Dimension	Questions to ask	Warning signs
The question	Was it the right question? Was it stated before the evidence was gathered?	The question changed to fit the answer
The method	Could the method have shown the answer was wrong? Did it fit the question?	No comparison, no control, no chance of a surprise
The evidence	Is it relevant, sufficient and accurately measured? What is missing?	Small or self-selected samples; unexplained exclusions
The reasoning	Do the conclusions follow? Were alternative explanations considered?	Confident claims that go beyond the data
The openness	Could someone else check or repeat the work?	Methods, data or decisions not recorded

Figure 2. Five dimensions for judging the quality of an inquiry independently of its answer. Table by ModQu.

The question. A good inquiry asks a question that matters to the decision at hand and states it before the evidence is gathered. If the question changed to fit the answer, the inquiry has shown only what it happened to find.

The method. A good method fits the question and could have produced an unwelcome answer. If no possible result would have been taken as evidence against the conclusion, the method did not test it.

The evidence. Good evidence is relevant, sufficient and accurately measured. A good inquiry also says what evidence is missing and how that limits the conclusions.

The reasoning. Good reasoning draws conclusions that follow from the evidence, considers alternative explanations and states uncertainty honestly.

The openness. A good inquiry records its methods, data and decisions so that others can check or repeat it. Openness does not guarantee quality, but its absence makes quality impossible to judge.

A hypothetical example

Two analysts are asked whether a new product will sell well. The first relies on a conversation with a few enthusiastic customers and predicts strong sales. The second runs a small, well-designed test with a representative sample, compares results against similar past launches and predicts moderate sales with a wide range of uncertainty. The product sells strongly.

Judged by outcome, the first analyst was right and the second was cautious. Judged by process, the first analyst was lucky and the second was sound. If the organisation rewards the first approach, it will rely on it for the next launch, and the next, until the luck runs out. If it reviews the inquiries themselves, it will see that the second analysis was the one worth repeating, and perhaps ask why its forecast was too low, which may reveal something useful about the market. A review of this kind takes little time. It needs only the original question, the plan and the forecast, recorded when the work was done, and someone willing to ask whether the method could have shown the forecast to be wrong.

Applying the principle

Some practical habits help organisations judge inquiries by their quality:

- **Review the inquiry before the answer is known.** Assess the question, method and evidence when the work is presented, not only after the results arrive.
- **Keep records.** Write down the question, the plan and the forecast at the time, so that later reviews are not distorted by hindsight.
- **Score many judgements.** Track forecasts and recommendations over time to separate skill from luck.
- **Review successes as well as failures.** Ask whether a good result came from a sound process, and do not assume it did. Regular debriefs that examine process have been shown to improve performance (Tannenbaum & Cerasoli, 2013).
- **Separate the reviewer from the result.** Where possible, ask someone who does not know the outcome to judge the quality of the inquiry.

Conclusion

The answer an inquiry produces is what most people care about, but it is a poor measure of the inquiry's quality. Luck, outcome bias and delay all separate good answers from good processes. Researchers in many fields have therefore developed criteria for judging inquiry on its own terms: whether the question was well chosen, whether the method could have shown the conclusion to be wrong, whether the evidence was adequate, whether the reasoning was sound and whether the work was open to checking. Organisations that apply these criteria, and that judge their inquiries before and independently of the results, will over time produce more reliable knowledge and make better decisions.

References

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology, 54*(4), 569–579. https://doi.org/10.1037/0022-3514.54.4.569

Dewey, J. (1938). *Logic: The theory of inquiry*. Henry Holt.

Duke, A. (2018). *Thinking in bets: Making smarter decisions when you don't have all the facts*. Portfolio.

Fischhoff, B. (1975). Hindsight ≠ foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance, 1*(3), 288–299. https://doi.org/10.1037/0096-1523.1.3.288

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460–465. https://doi.org/10.1511/2014.111.460

Goldman, A. I. (1979). What is justified belief? In G. S. Pappas (Ed.), *Justification and knowledge* (pp. 1–23). D. Reidel.

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ, 336*(7650), 924–926. https://doi.org/10.1136/bmj.39489.470347.AD

Higgins, J. P. T., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., Savović, J., Schulz, K. F., Weeks, L., & Sterne, J. A. C. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ, 343*, Article d5928. https://doi.org/10.1136/bmj.d5928

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), Article e124. https://doi.org/10.1371/journal.pmed.0020124

Kahneman, D., Lovallo, D., & Sibony, O. (2011). Before you make that big decision. *Harvard Business Review, 89*(6), 50–60.

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.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Sage.

Lovallo, D., & Sibony, O. (2010). The case for behavioral strategy. *McKinsey Quarterly*, (2), 30–43.

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press. https://doi.org/10.1017/9781107286184

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., & Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science, 25*(5), 1106–1115. https://doi.org/10.1177/0956797614524255

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

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

Popper, K. R. (1959). *The logic of scientific discovery*. Hutchinson.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Tannenbaum, S. I., & Cerasoli, C. P. (2013). Do team and individual debriefs enhance performance? A meta-analysis. *Human Factors, 55*(1), 231–245. https://doi.org/10.1177/0018720812448394

Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The art and science of prediction*. Crown.

Tracy, S. J. (2010). Qualitative quality: Eight "big-tent" criteria for excellent qualitative research. *Qualitative Inquiry, 16*(10), 837–851. https://doi.org/10.1177/1077800410383121