SACRED BOVINES

THE QUIXOTIC QUEST FOR OBJECTIVITY IN OBSERVATION

DOUGLAS ALLCHIN, DEPARTMENT EDITOR

What if the legendary character Don Quixote had been a scientist? Surely his quest would have been the noble pursuit of objectivity. Scientists endeavor to transcend mere opinion or individual interpretation. They strive for publicly confirmable facts. Accordingly, scientists appeal to empirical evidence, measurements, and observations—regarded as the bedrock for factual claims.

Yet, at the same time, ordinary humans can be fallible observers. Their interpretations can be skewed by prior expectations or personal desires. Historians, philosophers, and sociologists of science thus now typically contend that observations are "theory laden"—easily reflecting the researchers' assumptions. In the past, the ideal of science was expressed in the simple motto "I'll believe it when I see it!" Now, some cynics contend, an honest scientist might admit the ironic converse: "I'll *see* it when I *believe* it."

Are we inevitable puppets to our beliefs? To what degree are observations in science trustworthy? How else would we defend scientific claims? (How else would we resolve contentious facts in our society?) Most teachers, I think, endorse the conventional view—that scientists and their observations are inherently objective. And that this makes science privileged. Here I explore this revered view (this month's "Sacred Bovine"). Ultimately, I maintain, we are not as perfect as in the quixotic image. Yet science has developed tools to accommodate our cognitive flaws and to rescue science's claim to its much-vaunted objectivity.

○ Observer Bias & Blinding

Objectivity is a hallmark principle of our justice system too. Think of the allegorical figure holding aloft the scales of justice, blindfolded and impartial. Courts need trustworthy evidence to decide whether someone is culpable or innocent. For example, they rely on witnesses.

However, cognitive research has shown that observers' perceptions can be shaped and reshaped by personal experience and prejudices. Memories are vulnerable to suggestion too. Eyewitness testimony is—counterintuitively perhaps—among the *least* reliable in a courtroom (see the provocative volume by Loftus et al., 2019). That is, witnesses are susceptible to *observer bias*. We might, therefore, turn to forensic science and physical evidence—fingerprints, blood, DNA—as more secure.

But even here, observer bias can intrude. We know this because science has turned on itself, to investigate its own objectivity. Psychologists have tested forensic experts in historical crime scenarios. Their assessment of bullet and shoeprint evidence seemed pretty consistent. But when contextual information about a case was available, it could affect how they interpreted a crime scene, how they matched fingerprints, how they identified individuals from the DNA when a sample mixes DNA from multiple persons, how they interpreted bloodstain patterns, and how they assessed skin injuries, at least. Even what dog handlers believed about possible culprits could influence the behavior of their sniffer dogs (Colloff, 2018; Cooper & Meterko, 2019). What can be done to ensure justice?

Managing observer bias is standard now in modern medical research. To prevent judgment about a patient's condition being primed, the doctors are metaphorically blindfolded. They are not informed about who is receiving a new drug or treatment and who has been given an inert placebo. Bias is not possible, even unconsciously.

Such practices emerged over a century ago. One landmark study was done by Adolf Bingel in 1912–1913 at the City General Hospital in Brunswick, Germany (Tröhler, 2011). For decades, diphtheria had been a major scourge across Europe. Serum therapy (recognized in the very first Nobel Prize in Physiology or Medicine in 1901) had certainly improved the situation. Bingel acknowledged its efficacy but questioned whether it worked because of a specific antitoxin in the serum. Might the serum itself—any serum—be equally effective? By this time, the notion of controls for experimental comparison was widely appreciated (Sacred Bovines, March 2020). So, Bingel established two groups. Some patients received the conventional "antitoxin" serum, and others ordinary horse serum. To avoid inadvertently biasing his sample, he methodically assigned every other admitted patient to the alternate group.

Bingel was aware that given the controversial nature of his idea, the physicians' preconceptions posed a special danger. He reminded his readers that it is "extraordinarily difficult ¼ to evaluate the influences of therapy on disease unless they are obvious, as for example, the success of a surgical operation or cure of syphilis with mercury or Salvarsan. The therapeutic optimist very easily sees improvement, and the skeptic sees nothing." He thus wanted "to achieve an objective overall assessment," rather than the doctor's informal, possibly biased, "impressions." So, "to make the trial as objective as possible," he explained, "I have not relied on my own judgement alone but have sought the views of the [at least six] assistant physicians of the diphtheria ward, without informing them about the nature of the serum under test (namely the ordinary horse serum). Their judgement was thus completely without prejudice. I am keen to see my observations checked independently, and most warmly recommend this 'blind' method for the purpose" (Bingel, 1918, p. 288). Here, Bingel used the term still common today: *blinding*. That method gave stronger credence to Bingel's contentious conclusion: the theoretical claims of the Nobel Prize winner were mistaken. *Any* serum was effective.

The American Biology Teacher, Vol. 85, No. 2, pp. 122–124, ISSN 0002-7685, electronic ISSN 1938-4211. DOI: https://doi.org/10.1525/abt.2021.85.2.122.

○ Documenting Unconscious Bias

Documenting specific instances of observer bias can be difficult. However, one can gauge the magnitude of the general problem by bulk comparison of blinded and non-blinded observations. One such analysis looked at clinical studies about a range of medical treatments, from heart conditions to wounds to psychiatric disorders (Hróbjartsson et al., 2013; Hróbjartsson et al., 2014). In the non-blinded studies—the ones open to observer bias—the conclusions were (on average) more dramatic. Probabilities of benefit were 36% higher. Effect sizes increased by 68%. Similar discrepancies were found even for lab studies on animal models (Bello et al., 2014). Overall, blinded studies seemed to yield more modest results. Even among clinical trials with large, randomized samples, unwanted observer bias can intrude and yield misleading findings.

One might well imagine that observer bias would be limited to scientific studies where judgment is critical and where prior beliefs are strong. Not so. This method of comparing blinded and non-blinded studies has helped us probe that assumption (a further expression of this month's Sacred Bovine—that one may assume by default that a scientist's observations are immune to such influences).

For example, do ants recognize nestmates (their genetic kin)? According to the theory of kin selection, the behavior of an individual should tend to benefit its closest genetic relatives. So, this apparently simple question of insect behavior has significant implications for understanding evolutionary biology. A standard way to measure such kin-oriented behavior is to observe ants from the same versus different colonies meeting, and to tally the various types of encounters between them. To what degree do they exhibit aggressive behavior toward kin (nestmates) or toward "others"? Even with the relevant behaviors clearly defined, those assessments can be subtle, it turns out. Identifying instances of "mandible flaring" or "recoil" from a tactile encounter, for example, requires some experimenter judgment. In one recent meta-analysis, investigators found 156 experiments of nestmate versus non-kin behavior (van Wilgenburg & Elgar, 2013). Of those, 53 met the criteria for analysis of observer bias. Fifteen of those used blinded behavioral analysis. As was the case in the clinical studies, the results of the non-blinded studies tended to provide stronger evidence for the predominant theory. First, "aggression among nestmates was three times more likely to be reported in blinded than non-blinded experiments." Second, "the effect size—the differences between the level of aggression among nestmates and that among non-nestmates—in non-blind experiments was twice that of blind experiments." Here, blinded experiments seem to have escaped bias from theoretical expectations.

Another unlikely topic for observational error might be plant herbivory: namely, how much tree foliage do insects consume? One might envision a fairly straightforward task of sampling leaves and measuring the amount of loss—scan their surface area, weigh them, or count the proportion of leaves with damage. Or estimate defoliation visually, from photos of whole trees (and cross-check this method with some direct sampling). Simple measurements—manageable even by introductory students?

This topic, too, has been examined for evidence of observer bias—based on 42 publications of insect herbivory in Brazil (Kozlov et al., 2014). Again, blinded and non-blinded studies were compared. The plant damage differed by a factor of *five to ten*, depending on the methods used. Non-blinded studies reported significantly more damage than blinded studies. That is, they matched the widespread assumption that such rates are very high in the tropics. In addition, studies that focused on only one or a few species (1–3) found twice as much damage as those studying 10 or more species. Thus, the researcher's choice of individual species seems to have been a biasing factor. Perhaps one chooses a species because the damage is more noticeable (or "typical") to the observer who is seeking to measure it? Or the species is more prevalent, enabling easy sampling. But the selected species apparently did not fairly represent all species, and this error has led to misleading claims about insect herbivory in the tropics in general.

In a follow-up analysis (based on 125 publications), the same team identified other ways apparently insignificant choices seem to unconsciously bias such research: selection of study site; selection of timing (season and duration); and selection of individual branches or leaves to be sampled (Zvereva & Kozlov, 2019). Casual (technically, "haphazard") sampling can open the way to observer bias. In addition, primary authors who participated in the sampling or measurement, or others who knew where the samples had originated, inevitably inflated the magnitude of the results. The reviewers concluded sadly, "Our ecological and environmental knowledge is considerably biased due to an unconscious tendency of researchers to lend support for their hypotheses and expectations, which generally leads to overestimation of the effects under study." Blinding matters.

○ Pragmatic Horizons

These studies—of serum therapy, forensic analysis, clinical trials, ant behavior, and insect herbivory—document the widespread occurrence of unconscious observer bias in biology. Ironically, they equally indicate how blinding is effective in reducing its effects. Objectivity in science may be threatened by the infelicities of human observation, but it can also be salvaged by appropriate countermeasures. Accordingly, the custom of blinding—familiar to medical and psychological researchers for over a century now—is gradually informing more fields of science. (Note, too, its relevance to NGSS's third Scientific and Engineering Practice: Planning and Carrying Out Investigations.)

Observer bias is insidious, surely. Unconscious and easily hidden. It can severely threaten the quixotic ideal of objectivity in science. Yet turning a "blind eye" to such flaws only compounds the problem, allowing bias to fester at a yet deeper level. Fortunately, perhaps, while observer bias is unintentional, it can nonetheless be managed intentionally—through the strategy of blinding. In a society where facts are disputed, and allegations of prejudiced observations are rampant, such tools for reclaiming objectivity might well be more widely known—and perhaps fruitfully applied even by nonscientists.

References

Bello, S., Krogsbøll, L. T., Gruber, J., Zhao, Z. J., Fischer, D., & Hröbjartsson, A. (2014). Lack of blinding of outcome assessors in animal model experiments implies risk of observer bias. *Journal of Clinical Epidemiology*, *67*, 973–983.

Bingel, A. (1918). Über Behandlung der diphtherie mit gewöhnlichem pferdeserum. *Deutsches Archiv für Klinische Medizin*, *125*, 284–332.

Colloff, P. (2018, May 23 and 31). *Blood Will Tell* (Parts I and II). ProPublica. https://features.propublica.org/blood-spatter/blood-spatter-forensic-evidence-investigation/

Cooper, G. S., & Meterko, V. (2019). Cognitive bias research in forensic science: A systematic review. *Forensic Science International, 297* (April), 35–46.

Hróbjartsson, A., Thomsen, A. S., Emanuelsson, F., Tendal, B., Hilden, J., Boutron, I., Ravaud, P., & Brorson, S. (2013). Observer bias in randomized clinical trials with measurement scale outcomes: A systematic review of trials with both blinded and nonblinded assessors. *Canadian Medical Association Journal*, *185*, E201–211. https://doi.org/10.1503/cmaj.120744

Hróbjartsson, A., Thomsen, A. S., Emanuelsson, F., Tendal, B., Rasmussen, J. V., Hilden, J., Boutron, I., Ravaud, P., & Brorson, S. (2014). Observer bias in randomized clinical trials with time-to-event outcomes: Systematic review of trials with both blinded and non-blinded outcome assessors. *International Journal of Epidemiology*, *43*, 937–948.

Kozlov, M. V., Zverev, V., & Zvereva, E. L. (2014). Confirmation bias leads to overestimation of losses of woody plant foliage to insect herbivores in tropical regions. *PeerJ*, *2*, e709. https://doi.org/10.7717/peerj.709

Loftus, E. F., Doyle, J. M., Dysart, J. L., & Newirth, K. A. (2019). *Eyewitness Testimony: Civil and Criminal* (6th ed.). Lexis Law Publishing.

Tröhler, U. (2011). Adolf Bingel's blinded, controlled comparison of different anti-diphtheritic sera in 1918. *Journal of the Royal Society of Medicine, 104*, 302–305.

van Wilgenburg, E., & Elgar, M. A. (2013). Confirmation bias in studies of nestmate recognition: A cautionary note for research into the behaviour of animals. *PLOS ONE*, *8*, e53548. https://doi.org/10.1371/journal.pone.0053548

Zvereva, E. L., & Kozlov, M. V. (2019). Biases in studies of spatial patterns in insect herbivory. *Ecological Monographs*, *89*, e01361. https://doi.org/10.1002/ecm.1361