What irreproducible results mean for the law of scientific evidence

Jason M. Chin

In 2015, Brian Nosek and several collaborators reported the results of perhaps the most important scientific study of the year – it was a study they performed by copying the work of others. That is not as contradictory as it sounds, or at least it should not be.

While the conventional wisdom is that most scientific findings have been vetted and reproduced many times before they reach scientific consensus, the reality is that such rigour is exceedingly rare. Nosek sought to remedy this situation, wrangling up a group of 270 other researchers (collectively the Open Science Collaboration, or OSC) and attempting to redo 100 psychology experiments already published in leading peer-reviewed journals.[1] The question: Would these re-creations find the same results as the initial studies?

Prior to the OSC's endeavour, the conventional wisdom was that, yes, published science contains some false positives, but they are a small minority and quickly identified through a robust self-correction process. The OSC's results suggest that this conventional wisdom is wrong – only 36 percent of the studies in the sample were reproducible. This finding surprised even the alarmists, who long suggested the number was closer to 50 percent.[2] The OSC's results also garnered widespread media coverage, drawing attention to a problem that had been brewing for years.[3]

And while the OSC's study was in the context of psychological science, the picture is no rosier in other disciplines. For example, a large-scale replication attempt focused on cancer treatments was able to confirm the results of the original study just 11 percent of the time.[4] And, in December 2015, it was reported that a significant portion of research in clinical genetics (i.e., the type of research that provides for screening of genetic diseases) was unreliable.[5] Especially relevant to law, recent findings have put large swaths of the field of bite-mark forensics in doubt.[6]

Indeed, the legal community should heed the warnings flowing from the OSC's startling finding. Legal decisions often turn on scientific evidence. For example, courts use science to determine questions as varied as whether marijuana is addictive[7] and whether DNA links a suspect to a crime scene.[8] The problem may be even more immediate in law – whereas science self-corrects

(although, as we shall see, it takes longer than most expected), litigants rarely get more than one kick at the can.

Issues with scientific evidence rose to prominence in Canada with concerns over the practice of pediatric forensic pathology and the subsequent report prepared by a commission of inquiry led by Justice Stephen Goudge.[9] The report described serious flaws in the practice of forensic pathology, including with the skill and objectivity of expert witnesses.

The OSC reported a more pervasive problem, going beyond the practice of experts in a field to the basic science that girds their opinions. In other words, the safeguards built into the very *process* of scientific inquiry have long been inadequate to weed out erroneous findings. That process is changing (for the better) and legal actors must be aware of these changes.

In what follows, I first provide a brief review of the legal standard for the admission of scientific evidence. Then, I review several reasons why science has proven unreliable. These reasons inform the steps legal actors can take to effectively screen out unreliable science. In short, they should demand science's best rather that what has been generally accepted. I conclude with some tangible recommendations for the justice system.

The admission of scientific evidence

The admission of scientific expert evidence in Canada,[10] US federal courts, and most US state courts is governed by the landmark US Supreme Court decision *Daubert v. Merrell Dow*.[11] *Daubert* provides four questions trial judges should consider when determining whether proffered scientific evidence is admissible: (1) Is it generally accepted? (2) Has it been peer reviewed and published? (3) Has it been tested? (4) What is the error rate of the finding or method? In Canada, this inquiry accompanies the general test for the admission of expert evidence found in *R v. Mohan*.[12]

Daubert has attracted a great deal of scrutiny from both academics and practitioners.[13] The test's critics focus on the enhanced gatekeeper role played by judges under its framework. They argue that trial judges, who frequently do not come from a scientific background, do not possess the

scientific fluency to apply *Daubert*. Indeed, some empirical research shows that judges fail to spot key flaws in scientific studies presented to them.[14]

In Canada, there appears to be more agreement that an enhanced gatekeeper role under *Daubert* is desirable. Indeed, this is one of the primary themes arising from the Goudge Report.[15] And, in a 2015 Ontario Court of Appeal decision, Chief Justice George Strathy cited that report in enunciating Ontario's strong position regarding judicial gatekeeping: "[T]here has been growing recognition of the responsibility of the trial judge to exercise a more robust gatekeeper role in the admission of expert evidence."[16]

To understand fully what science's reproducibility problem means for courts, more background is required. The following section reviews the sources of the problem, which all flow from a flawed incentive system.

Sources of the failure to replicate

Science's incentive structure values publishability over truth.[17] When one contemplates a research project, the goal line is almost inevitably publication. It is the metric by which scientists are measured. Indeed, publishing has been described as "the very heart of modern academic science – at levels ranging from the epistemic certification of scientific thought to the more personal labyrinths of job security, quality of life and self-esteem."[18]

The problem is that the qualities that make findings publishable are not the same ones that make them true. In fact, the competition to publish often encourages shortcuts and questionable practices that further bias the body of published research. This disconnect between publishability and truth manifests in several ways: (1) research is rarely independently replicated; (2) studies failing to find a statistically significant effect (i.e., negative results) are almost never published; and (3) researchers sometimes misbehave. These issues are reviewed below.

Journals value novelty, not replication

Replication is commonly referred to as the *sine non qua* of science,[19] and for good reason. Direct replication by an independent party substantially increases the confidence in a result.[20] In

particular, direct replication helps confirm that the finding was not a statistical artifact and reduces the possibility that the experimenter deliberately or subconsciously affected the result.

Despite unanimous acceptance of the importance of replication, leading journals rarely publish direct replications of previous research. The reason for this is simply that journals value novelty over truth. In other words, they prefer to publish findings that are exciting, novel and counterintuitive, rather than replications of previous work.[21]

Both anecdotal accounts and empirical studies demonstrate the preference for novelty over replication. For example, surveys of editors find most agree that replications are a waste of valuable journal space.[22] Further, a 2012 study found that only 1.07 percent of psychological studies were replicated.[23]

Journals value positive results rather than negative ones

Similar to the problem with publishing replications, researchers also struggle to find an outlet for studies that find a negative result (i.e., that a treatment had no discernible effect). As a result, the publicly available body of evidence is positively biased, giving the impression that a finding is robust, when in fact there are many unpublished studies that show no effect. This bias against negative results is empirically verifiable: A 2012 study estimated that 90 percent of published results across all scientific fields are positive results.[24]

Questionable research practices

The incentive to publish creates a conflict between the personal interests of the scientist and the goal of knowledge accumulation (i.e., accuracy).[25] When personal motivations supersede accuracy, researchers may engage in questionable research practices (QRPs). QRPs are techniques used by researchers to strategically (but artificially) inflate the persuasiveness of their findings. These practices do not rise to the level of academic fraud, but undoubtedly they bias the state of knowledge in a field.[26]

More than 60 percent of researchers responding to an anonymous survey admitted to using one particular QRP – selectively reporting the measurements that confirmed the study's hypothesis

and refraining from reporting those that did not.[27] Although this practice may seem relatively benign, it skews error rates by increasing the chance that researchers will find a statistically significant result.

Although use of QRPs represents misbehaviour by scientists, journals share some of the blame. In particular, they have failed to scrutinize adequately the findings they publish. *Nature,* a leading scientific journal, expressly admitted its culpability in the replicability crisis:

> The problems arise in laboratories, but journals such as this one compound them when they fail to exert sufficient scrutiny over the results they publish and when they do not publish enough information for other researchers to assess results properly.[28]

The myth of self-correction

Science's claim to fame has long been that it self-corrects – erroneous findings are quickly and efficiently scrutinized and, if necessary, weeded out by other scientists. As Nosek described the problem:

> If a claim is wrong, eventually new evidence will accumulate to show that it is wrong and scientific understanding of the phenomenon will change. This is part of the promise of science – following the evidence where it leads, even if it is counter to present beliefs. ... We do believe that self-correction occurs. Our problem is with the word "eventually." *The myth of self-correction* is recognition that once published, there is no systemic ethic of confirming or disconfirming the validity of an effect. False effects can remain for decades, slowly fading or continuing to inspire and influence new research. ... Further, even when it becomes known that an effect is false, retraction of the original result is very rare.[29]

Fortunately, and thanks to scientists who have spent the past decade examining the problem, solutions are falling into place. Armed with these lessons, the law of scientific evidence can develop to ensure reproducible science takes centre stage in legal disputes.

What can the legal system do about it?

Daubert provides a useful framework for courts to evaluate science. But the *indicia* of good science *Daubert* directs judges to consider (i.e., peer review, testing and error rates) have themselves failed to track best practices. For example, the OSC's results suggest that the fact that a study was peer reviewed carries little weight. It is not the mere act of peer review that matters, but how thorough that process was. I will now review the *Daubert* factors and note how legal actors can employ them in the context of the OSC's findings, to ensure that bad science does not have an impact on legal decisions.

Peer review and publication

The OSC's results suggest that peer review and publication, in and of itself, is a poor *indicium* of scientific reliability because only a minority of such studies prove replicable. In fact, a poorly carried out peer review process may do no more than increase the superficial believability of a study without providing any guarantees of reliability. Recall that some journals have expressly admitted culpability, stating that they failed to scrutinize sufficiently the results they published or did not publish enough information for others to scrutinize them. By placing weight on peer review for its own sake, *Daubert* may also distract judges from unpublished work containing valuable data.

Several initiatives can help to ensure that *Daubert's* peer review component is effectively used. For instance, the Goudge Report recommended judicial training programs, which appear effective in several US jurisdictions.[30] These programs, as well as guides prepared for use by courts, could draw attention to the levels of scrutiny provided by various journals. (Nosek's organization, for example, awards badges to publications that adopt more open practices, such as by making their data publicly accessible and thus open to scrutiny.[31])

Informed lawyers would do well to cross-examine experts not only on whether the findings they rely on have been peer reviewed and published, but also on *how* that process was carried out. Are the authors' data available to the public? Are the instruments they used to make measurements available and accessible? Compelling proprietary reasons may exist for non-disclosure, but that is often not the case.

Even when expert evidence passes the threshold for admissibility, a battle continues over which expert's evidence will be accorded greater weight. In such cases, one relevant factor should be the quality of peer review the scientific evidence has endured. Peer review is not always rigorous and, all else being equal, scientific evidence that has withstood a higher level of scrutiny should receive the greater weight.

Ontario courts have been especially vigilant to the importance of properly performed peer review and publication. For instance, in *R. v. Truscott*,[32] a five-justice panel of the Ontario Court of Appeal was extremely critical of the Crown's expert witness whose opinion rested on little more than anecdotal experience. Instead, the court endorsed an approach centred on a "*critical analysis* of peer-reviewed literature."[33] Ontario advocates may wish to refer to *Truscott* in arguing that recent insights into the reliability of science demonstrate that the bar for critical analysis of peer-reviewed evidence is high.

In addition, expert witnesses in Ontario are now required to sign an acknowledgement of their duties. As part of this process, scientific experts could be asked to acknowledge that they attempted to canvass fully the relevant subject area and sought unpublished research that either confirms or casts doubt on the science at issue. Experts would need to contact scientists known to be active in a field and request their unpublished data. (A major contributor to the replicability crisis is that high-quality research with negative findings, although available through informal channels, is seldom published. Likewise, court-appointed experts in cases involving complicated scientific issues may be directed to undertake a similar endeavour.[34]

Testing and error rates

Scientists use a variety of tactics to artificially inflate the persuasiveness of their results. Although these testing practices were once generally accepted, there is now no doubt that they have introduced bias into several bodies of scientific findings. Fortunately, it is relatively easy for courts and lawyers to be on the lookout for QRPs.

Scientific journals increasingly require researchers to adopt open methods. For example, some require researchers to "pre-register" their studies (or at least make note of those who do so). Pre-registration means that the researcher used one of several existing web-based databases to pre-

record the experiment's parameters and helps to ensure that the researcher did not engage in QRPs. For example, because researchers must publicly pre-record the measurements they are going to make, pre-registration catches the QRP of failing to report them.

Absent a compelling reason, most modern scientific research should be pre-registered. Well-informed litigators would also be wise to inquire about pre-registration during cross-examination, when appropriate. And, once again, even if the scientific evidence is admitted, the presence or absence of pre-registration should inform the evidence's weight.

Large-scale replication attempts, like the OSC's, are exceedingly rare. While it is unsettling that so many of the studies in its sample proved irreproducible, 36 percent were directly replicated by researchers acting independently of those who first made the finding. This type of direct replication, when it does occur, allows scientists to hone in on an accurate estimate of the error rate associated with a finding or technique. There are few situations, if any, in which non-replicated research should be admitted to influence the ultimate decision on the merits. Adjudicators should therefore adopt a skeptical attitude toward unreplicated evidence, questioning whether the extant evidence has been sufficiently tested and represents an accurate representation of the error rate of the finding or technique.

Conclusion

The meta-scientific insights described herein have not gone unnoticed. Nosek's own organization has received millions of dollars of support and now provides a host of tools for pre-registration of studies and support throughout the life of a scientific experiment.[35] Journals have also taken note and are enhancing their review processes. The legal community should take heed of these changes.

Although the situation sounds dire, science is still the best way to accumulate knowledge about the world. There are, however, more – and less – accurate ways of accomplishing this goal. We now know more about the dangers of the less accurate ways, and we have improved guidance on the more accurate ways. By attending to the methods that foster reproducibility, courts can better ensure that legal disputes are adjudicated accurately.

Notes

1. Open Science Collaboration, "Estimating the Reproducibility of Psychological Science" (2015) 349 Science 943.
2. John PA Ioannidis, "Why Most Published Findings Are False" 2:8 PLoS Med e124.
3. Ed Yong, "How Reliable Are Psychology Studies?" *The Atlantic* (27 August 2015), online: <theatlantic.com/science/archive/2015/08/psychology-studies-reliability-reproducability-nosek/402466>; Joel Echenbach, "Many Scientific Studies Can't Be Replicated: That's a Problem," *The Washington Post* (27 August 2015), online: <washingtonpost.com/news/speaking-of-science/wp/2015/08/27/trouble-in-science-massive-effort-to-reproduce-100-experimental-results-succeeds-only-36-times>; Benedict Carey, "Many Psychology Findings Not as Strong as Claimed, Study Says," *The New York Times* (27 August 2015), online: <nytimes.com/2015/08/28/science/many-social-science-findings-not-as-strong-as-claimed-study-says.html>.
4. C Glenn Begley & Lee M Ellis, "Drug Development: Raise Standards for Preclinical Cancer Research" 483 Nature 531.
5. Ed Yong, "Clinical Genetics Has a Big Problem That's Affecting People's Lives," *The Atlantic* (16 December 2015), online: <theatlantic.com/science/archive/2015/12/why-human-genetics-research-is-full-of-costly-mistakes/420693>.
6. Erik Eckholm, "Lives in Balance, Texas Leads Scrutiny of Bite-Mark Forensics" *The New York Times* (12 December 2015), online: <nytimes.com/2015/12/13/us/lives-in-balance-texas-leads-scrutiny-of-bite-mark-forensics.html>.
7. *R v Smith*, 2015 SCC 34, [2015] 2 SCR 602.
8. *R v B(SA)*, 2003 SCC 60, [2003] 2 SCR 678.
9. Ontario, *Inquiry into Pediatric Forensic Pathology in Ontario: Report* (Toronto: Ontario Ministry of the Attorney General, 2008) (Commissioner Stephen T Goudge);online: < attorneygeneral.jus.gov.on.ca/inquiries/goudge/index.html> [Goudge].
10. *R v J(LJ)*, 2000 SCC 51 at para 33, [2000] 2 SCR 600 [*JLJ*].
11. 509 US 579–595 (1993).
12. [1994] 2 SCR 9, 1994 CarswellOnt 1155; *JLJ*, *supra* note 10 at paras 25–62.
13. See e.g. Ken J Chesebro, "Taking Daubert's 'Focus' Seriously: The Methodology/Conclusion Distinction" (1994) 15 Cardozo L Rev 1745.
14. Margaret Bull Kovera & Bradley D McAuliff, "The Effects of Peer Review and Evidence Quality on Judge Evaluations of Psychological Science" (2000) 85 J Applied Psychology 574.
15. Goudge, *supra* note 9.
16. *Meady v Greyhound Canada Transportation Corp*, 2015 ONCA 6 at para 37, 2015 CarswellOnt 46.
17. Brian A Nosek, Jeffrey R Spies & Matt Motyl, "Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth over Publishability" (2012) 7:6 Perspectives on Psychological Science 615 [Utopia].
18. Michael J Mahoney "Open Exchange and Epistemic Process" (1985) 40 Am Psychologist 29.

19. David C Funder et al, “Improving the Dependability of Research in Personality and Social Psychology: Recommendations for Review and Educational Practice” (2014) 18 Personality and Social Psychology Rev 3 at 8.
20. Harold Pashler & Christine R Harris, “Is the Replicability Crisis Overblown? Three Arguments Examined” (2012) 7 Perspectives on Psychological Science 531 at 534.
21. Utopia, *supra* note 17 at 617.
22. James W Neuliep & Rick Crandall, “Editorial Bias against Replication Research” (1990) 5 J Social Behavior and Personality 85.
23. Matthew C Makel, Jonathan A Plucker & Boyd Hegarty, “Replications in Psychological Research: How Often Do They Really Occur?” (2012) 7 Perspectives on Psychological Science 537.
24. Daniele Fanielli, “Negative Results Are Disappearing from Most Disciplines and Countries” (2012) 90 Scientometrics at 891.
25. Utopia, *supra* note 17 at 616.
26. Joseph P Simmons, Leif D Nelson & Uri Simonsohn, “False Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant” (2011) 22 Psychological Science 1359.
27. Leslie K John, George Loewenstein & Drazen Prelec, “Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling” (2012) 23:5 Psychological Science 524 at 525.
28. Editorial “Announcement: Reducing Our Irreproducibility” (2013) 496 Nature 398; see also Eric Eich “Business Not as Usual” (2014) 25 Psychological Science 3.
29. Utopia, *supra* note 17 at 619.
30. GT Harrell Jr et al, “Deployment of Science and Technology Trained Judges: Settling on a Plan” (Washington DC: Advanced Science and Technology Adjudication Center, 2009).
31. Online: <osf.io/tvyxz/wiki/home>.
32. 2007 ONCA 575, 2007 CarswellOnt 5305.
33. *Ibid* at para 169 [emphasis added].
34. Goudge, *supra* note 9 at 506.
35. Online: <osf.io>.