R v Bingley and the importance of scientifically-guided legal analysis*

Jason M. Chin & Helena Likwornik**

*Cite as:

Jason M Chin & Helena Likwornik, “*R v Bingley* and the importance of scientifically-guided legal analysis” (forthcoming 2017) *Queen’s Law Journal*.

**Jason M. Chin, Lecturer, TC Beirne School of Law, University of Queensland. Helena Likwornik – the views expressed by Helena Likwornik represent only her personal views and are not intended to reflect the views of any institution(s) with which she is associated. We thank Jocelyn Bosse, Caitlin Goss, Hamish McDonald, and three anonymous reviewers for their thoughtful and perspicacious feedback. This research was supported by a start-up grant from the University of Queensland awarded to the first author.

Abstract

In *R v Bingley*, the Supreme Court considered a controversial subjective methodology used by police officers trained as drug recognition experts (DREs) pursuant to the *Criminal Code*. At issue was the admissibility of these experts' evidence. A 5-2 majority held that Parliament conclusively established the reliability the DRE program's methodology and the DRE's qualifications to perform that methodology, and thus trial judges may not exclude DREs for those reasons. *Bingley* is problematic on multiple fronts. Most fundamentally, the Majority's statutory interpretation was insensitive to the science behind the drug recognition program. Their analysis put this subjective methodology on the same footing as objective forms of evidence, like breathalyser analysis, where human judgment and bias play almost no role. More broadly, the Majority's decision comes in light of recent findings that several forensic scientific disciplines are not as reliable as they purport to be. The Majority's reasoning seemed largely driven by concerns about judicial economy, and in particular the worry that evaluating DREs would take too much court time. In response, we provide a more scientifically rigorous but less time-consuming way for trial judges to scrutinize DREs.

Part I. Introduction

The unmistakable overall trend of the jurisprudence, however, has been to tighten the admissibility requirements and to enhance the judge's gatekeeping role.¹

The Supreme Court, reflecting on the past 20 years of expert evidence jurisprudence, made the above statement in 2015. It was an accurate assessment. Canadian high courts have been increasingly concerned with the dangers of expert evidence, especially that evidence comes from scientists cloaked in the “mystique of science”² and police officers whose authority can provide a “superficial attractiveness”.³ As a result, these courts have reminded trial judges of their duty to vet expert evidence before it reaches the trier of fact and have accordingly empowered them to do so.⁴ The Supreme Court's recent decision in *R v Bingley*⁵ tested the extent of this trend towards expanding the trial judge's gatekeeper role and seems to have identified its limit. We will suggest that it is not a scientifically sound limit.

The stakes in *Bingley* go well beyond the case itself. It arises in a time of upheaval within the forensic sciences that has seen peak scientific bodies, including the United States National

¹ *White Burgess Langille Inman v Abbott and Haliburton Co*, 2015 SCC 23 at paras 20, [2015] 2 SCR 182 [*White Burgess*].

² *R v D(D)*, 2000 SCC 43 at para 2, 191 DLR (4th) 60 [DD]; *R v Mohan*, [1994] 2 SCR 9 at para 23, 114 DLR (4th) 419 [*Mohan*].

³ *R v Sekhon*, 2014 SCC 15 at para 50, [2014] 1 SCR 272 [*Sekhon*]. For a review of this jurisprudence, see: Lisa Dufrainmont, “New Challenges for the Gatekeeper: The Evolving Law on Expert Evidence in Criminal Cases” (2011-2012) 58 Criminal Law Quarterly 531-558.

⁴ Note, however, that several commentators have questioned whether courts regularly make use of these powers. See: Gary Edmond & Kent Roach, “A Contextual Approach to the Admissibility of the State's Forensic Science and Medical Evidence” (2011) 61 UTLJ 343 [Contextual Approach] and Gary Edmond & Emma Cunliffe, “Cinderella Story? The Social Production of a Forensic ‘Science’” (2016) 106:2 The Journal of Criminal Law and Criminology 219-273. Similarly, apparently expert opinion evidence is sometimes characterized as lay opinion, thus skirting the expert gatekeeper: Gary Edmond & Mehera San Roque, “Quasi-Justice: Ad Hoc Expertise and Identification Evidence” (2009) 33 Crim LJ 8; Edward J Imwinkelried, “Distinguishing Lay From Expert Opinion: the Need to Focus on the Epistemological Differences Between the Reasoning Process Used by Lay and Expert Witnesses” (2015) 68 SMU L Rev 73; and, Jason M Chin et al, “Drawing the Line Between Lay and Expert Opinion Evidence” (2017) 63:1 McGill Law Journal.

⁵ 2017 SCC 12, 135 WCB (2d) 356 [*Bingley*].

Academy of Sciences (“NAS”), call into question the reliability of several long established forensic disciplines.⁶ Similar issues abound in the basic sciences.⁷ Simply put, this is not the time to punctuate a two decades long trend of empowering trial judges to question science. Trial courts need the flexibility to question a body of knowledge that is in flux.

The science at issue in *Bingley* was a standardized but subjective method employed by specially-trained police officers, Drug Recognition Experts (DREs), to determine whether a suspect is drug-impaired (collectively, the Drug Recognition Expert Program or DREP). DREs perform a 12-step assessment pursuant to sections 254 and 254.1 of the *Criminal Code*⁸ and the *Evaluation of Impaired Operation (Drug and Alcohol) Regulations* (the Regulations).⁹ This legislative scheme came into force in 2008. It provides a sister regime for laws in place to detect and enforce alcohol-impaired driving (e.g., breathalysers to measure blood alcohol concentration). Under the DREP, a police officer with a reasonable suspicion that a motorist is under the influence of a drug may demand that motorist be examined by a DRE. Pursuant to the Regulations, DREs must have taken courses with and been accredited by the International Association of Chiefs of Police (IACP), a non-governmental organization based in the United States.

⁶ National Research Council, *Strengthening Forensic Science in the United States: A Path Forward* (Washington, DC: National Academies Press, 2009) [NAS Report].

⁷ D Stephen Lindsay, “Replication in Psychological Science” (2015) 26:12 *Psychological Science* 1827-1832; Marcus Munafò et al, “A manifesto for reproducible science” (2017) 1 *Nature Human Behaviour* 1-9; Denes Szucs & John P A Ioannidis, “Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature” (2017) 15:3 *PLoS Biol* e2000797; Florian Prinz, Thomas Schlange & Khusru Asadullah, “Believe it or not: how much can we rely on published data on potential drug targets?” (2011) *Nature Reviews Drug Discovery*.

⁸ *Criminal Code*, RSC 1985, c C-46 [*Criminal Code*]. Note that this program has recently received considerable negative media attention in the U.S. for alleged false positives, see: Brendan Keefe & Michael King, “The Drug Whisperer: Drivers arrested while stone cold sober” 11 *Alive NBC News* (2017) online: <http://www.11alive.com/news/investigations/the-drug-whisperer/437061710>.

⁹ SOR/2008-196 [*Regulations*]; *Bingley*, *supra* note 5 at para 4.

The Majority and Dissent disagreed as to whether the *Criminal Code* had established that a properly administered 12-step DREP assessment was sufficiently reliable to serve as evidence of impairment in court, or alternatively whether it was still open to scrutiny by the trial judge. The Dissent agreed with the accused: the text of the DREP provisions contrasted with those surrounding breathalysers and thus they should not be treated the same in court.¹⁰ The breathalyser provisions render them “conclusive proof”¹¹ of intoxication, whereas there is no such language with regard to DREs’ evidence. The Majority held otherwise: While a DRE’s conclusions may be scrutinized on limited grounds, the 12-step process itself is irrebuttably valid and reliable.¹²

We propose that the statutory interpretation in *Bingley* should have been guided by the science. It is not just that the words in the breathalyser and DREP provisions are different; *so is the science*. Unlike breathalyser results, the DREP is subjective and thus there may be vast differences between DREs in their ability and skill in applying the methodology. Further, errors in application are more frequent in subjective methodologies.

In the remainder of this Part, we will review the facts of *Bingley*, and the Majority and Dissent’s reasons. Part II situates the DREP within the forensic sciences. Building on that foundation, Part III argues that Parliament did not intend for the DREP to elude the trial judge’s gatekeeper role because it is a subjective methodology and Parliament focused on its power to identify the presence of drugs as opposed to its sizable false positive rate. In Part IV, we discuss *Bingley*’s ramifications for expert evidence law writ large. And, finally, in Part V, we offer a

¹⁰ *Bingley*, *supra* note 5 at para 54; Appellant’s Factum at the Supreme Court of Canada, file number 36610, available online at < http://www.scc-csc.ca/WebDocuments-DocumentsWeb/36610/FM010_Appellant_Carson-Bingley.pdf > at paras 44-59 [Bingley’s Factum].

¹¹ *Criminal Code*, *supra* note 8 at s 258(1)(c).

¹² *Bingley*, *supra* note 5 at paras 27-28, 30, 32.

solution to manage DRE evidence in court that is scientifically rigorous but also judicially economical.

The accused in *Bingley* was observed driving erratically. Pursuant to the DREP, he was evaluated by Constable Jellinek, a certified DRE. DREs, including Constable Jellinek, employ a standardized subjective methodology to determine whether a target is impaired by drugs.¹³ The methodology comprises 12 steps, including judgments of physical coordination, balance, pupil size, eye gaze, muscle tone and pulse. If the evaluation affords “reasonable grounds” to believe the suspect’s ability to drive is impaired by a drug, the DRE may then demand a sample of blood, saliva or urine as confirmatory evidence.¹⁴ The Constable concluded that Mr. Bingley was under the influence of a drug.¹⁵ A urinalysis revealed the presence of THC, but was inconclusive as to when that exposure occurred.¹⁶ At trial, Bingley successfully challenged the reliability of Constable Jellinek’s subjective evaluation.¹⁷

The case eventually made its way to the Supreme Court¹⁸ where Chief Justice McLachlin, writing for the Majority, held that the evidence of a DRE should be admitted without a *voir*

¹³ Douglas J Beirness, Jacques LeCavalier & Deanna Singhal, “Evaluation of the Drug Evaluation and Classification Program: A Critical Review of the Evidence” (2007) 8 Traffic Injury Prevention 368-276 at 369 [Beirness 2007]; *Bingley*, *supra* note 5 at paras 8-10.

¹⁴ *Criminal Code*, *supra* note 8 at 254(3.4).

¹⁵ *Bingley*, *supra* note 5 at para 2.

¹⁶ *R v Bingley*, 2015 ONCA 439 at para 15, 126 OR (3d) 525 [*Bingley* ONCA].

¹⁷ *Ibid* at para 24.

¹⁸ The procedural history is lengthy but can be summarized as follows: Bingley was acquitted in two trials and in both cases the Crown successfully obtained new trials by way of summary conviction appeal. The first acquittal was overturned because the trial judge failed to consider the “cumulative effect of the evidence”. At the second trial, the trial judge found that Officer Jellinek’s opinion that Bingley’s driving was impaired due to drugs was inadmissible because it lacked a scientific basis. He was once again acquitted. This resulted in a second summary conviction appeal, in which the judge ordered a new trial because Officer Jellinek should not have been subjected to a *Mohan voir dire*. Bingley appealed the second summary conviction order to the Court of Appeal for Ontario, which dismissed the appeal. The appellate court found that both the object and text of section 243(3.1), which provides that the DRE is to “determine” whether a driver is drug-impaired, obviates a *voir dire*. A new trial was ordered, and Bingley appealed to the Supreme Court. See *ibid* at paras 6-31, 44-56.

dire.¹⁹ The Majority’s analysis of the text, context and legislative intent²⁰ of the DREP provisions concluded that the qualifications of DREs had been “conclusively and irrebuttably established by Parliament”²¹ and the associated scientific validity “comes from statutory framework itself.”²² It was common ground at the Supreme Court that the bodily fluid analysis component of the DREP was admissible.

Justice Karakatsanis, writing for herself and Justice Gascon, disagreed; they would have held that DRE evidence is open to scrutiny by trial judges.²³ More specifically, they could find no wording in the DREP provisions nor their legislative history to indicate Parliament intended a DRE’s evidence should evade the trial judge’s gatekeeping function. As noted above, the DREP provisions were textually distinct from those covering breathalyser results, the latter being admissible due to express language in the *Criminal Code* rendering them “conclusive proof” of the “concentration of alcohol in the accused’s blood”.²⁴ Rather, Parliament envisioned the DREP as an investigatory tool and thus approved its reliability for that purpose.²⁵ Accordingly, “courts retain discretion to require – through evidence or precedent – confirmation that the science behind DRE evaluations meet the necessary level of reliability before admitting the evidence at trial.”²⁶

Understanding the divergence between the Majority and Dissent requires some background into the science of the DREP. In the following section, we will situate the DREP within forensic science, a field that has received recent criticism for failure to provide empirical

¹⁹ *Bingley*, *supra* note 5 at paras 13-33.

²⁰ *Ibid* at paras 23-28.

²¹ *Ibid* at para 27.

²² *Ibid* at para 24.

²³ *Ibid* at paras 56-69.

²⁴ *Ibid* at paras 44-59; *Criminal Code*, *supra* note 8 at s 258(1)(c).

²⁵ *Ibid* at para 40.

²⁶ *Ibid*.

support for its claims. The DREP is a subjective forensic science; unlike more objective methods, its conclusions are profoundly affected by human judgment.

Part II. The DREP as a Subjective Forensic Science

The DREP is a forensic scientific method. Forensic science (i.e., science as it applies to criminal law) has been deeply criticized over the past several decades for the use of unreliable methods and the failure of organizations of practitioners to self-police.²⁷ These criticisms have come from both scientific bodies and the legal community. In 1998, Justice Fred Kaufman reported the results of his investigation into the wrongful conviction of Guy Paul Morin.²⁸ He found that flawed forensic science, particularly hair and fibre identification, contributed to the jury's decision to convict Morin.²⁹ Ten years later, Justice Stephen Goudge detailed numerous grave failures in the practice of forensic pediatric pathology in Ontario.³⁰ Both jurists recommended trial judges play a stronger role as gatekeepers of forensic science.³¹ In other words, they should refuse to admit such evidence when it lacks a valid scientific foundation.

The Supreme Court of Canada responded to these miscarriages of justice by changing expert opinion evidence law. The Court expanded the role of trial judges, directing them to review the validity of scientific evidence (i.e., how it was tested, its error rate, whether it was peer reviewed and published, and whether it is generally accepted) when that evidence is novel

²⁷ Barack Obama, "The President's Role in Advancing Criminal Justice Reform" (2017) 130:3 Harv L Rev 811 at 860-862.

²⁸ Fred Kaufman, *The Commission on Proceedings Involving Guy Paul Morin* (Toronto: Ontario Ministry of the Attorney General, 1998) [Kaufman].

²⁹ *Ibid* at 83.

³⁰ Stephen T Goudge, *Inquiry into pediatric forensic pathology in Ontario* (2008) Retrieved from <<http://www.attorneygeneral.jus.gov.on.ca/inquiries/goudge/report/index.html>> [Goudge].

³¹ Kaufman, *supra* note 28 at 324: "In my respectful view, however, it is appropriate for trial judges to undertake a far more critical analysis of the admissibility of this kind of evidence. My own view is that hair comparison evidence of the kind introduced in the *Morin* case should rarely be admitted for inclusionary purposes."; Goudge, *ibid* at 496-502.

science.³² Later in *R v Trochym* and *White Burgess Langille Inman v Abbott and Haliburton Co.*, the Court added that “contested science” should also receive this scrutiny.³³

Since these inquiries into specific convictions and practices, empirical research has confirmed that invalid forensic science is present in about 60% of wrongful convictions.³⁴ These sobering findings have inspired peak scientific bodies to investigate the shortcomings in forensic scientific theory and practice that may be accountable. For example, in 2009, the NAS published a report, led by prominent judges and scientists, describing a three-year study into the state of forensic science (the “NAS Report”).³⁵ This report was remarkably critical: “the forensic science system exhibits serious shortcomings in capacity and quality; yet the courts continue to rely on forensic evidence without fully addressing the limitations of different forensic science disciplines.”³⁶ The Report provided a framework for change, including recommending increased funding to the forensic sciences and further research.

In response to the NAS Report, President Obama directed his President’s Council of Advisors on Science and Technology (PCAST) to conduct additional research on how forensic science should be improved. PCAST published a report in 2016 (the “PCAST Report”).³⁷ The PCAST Report found that deep flaws still exist in several forensic methodologies, both with

³² *R v J(J-L)*, 2000 SCC 51 at paras 33-36, [2000] 2 SCR 600 [JLJ]; *Daubert v Merrell Dow Pharmaceuticals Inc.*, 509 US 579 (1993), 113 S Ct 2786 [Daubert]

³³ *R v Trochym*, 2007 SCC 6 at paras 32-36, [2007] 1 SCR 239 [Trochym]; *White Burgess*, *supra* note 1 at para 23.

³⁴ Brandon L Garrett & Peter J Neufeld, “Invalid Forensic Science Testimony and Wrongful Convictions” (2009) 95:1 Va L Rev 1.

³⁵ NAS Report, *supra* note 6. For two cogent discussions of the NAS Report, see: Jennifer L Mnookin, “The courts, the NAS, and the future of forensic science” (2010) 75:4 Brooklyn Law Review 1209-1275 and David H Kaye, “Probability, individuation, and uniqueness in forensic science evidence: listening to the academics” (2010) 75:4 Brooklyn Law Review 1163-1185.

³⁶ NAS Report, *ibid* at 53.

³⁷ United States, President’s Council of Advisors on Science and Technology, *Forensic Science in Criminal Court: Ensuring Scientific Validity of Feature-Comparison Methods* (Washington DC: Executive Office of the President, 2016) [PCAST Report]; and for additional reviews of the forensic sciences, see: Contextual Approach, *supra* note 4 and Gary Edmond et al, “Contextual Bias and Cross-contamination in the Forensic Sciences: the Corrosive Implications for Investigations, Plea Bargains, Trials and Appeals” (2014) 13 Law, Prob & Ris 1 [Contextual Bias].

respect to their foundational validity (can the method accurately make the identification it purports to make?) and their validity-as-applied (can practitioners consistently and accurately apply the methodology in the field?).

Particularly relevant to the DREP, the PCAST Report found that deficiencies in validity ran deeper for subjective methods, which rely on human judgment to a greater extent than objective methods. For example, bitemark analysis, like the DREP, is highly subjective, relying on a practitioner's comparisons of a found bitemark to known dentition.³⁸ On the other hand, DNA analysis, which relies on automated comparisons performed by computer algorithms, is predominantly objective.

Foundational validity requires (1) consistent and reproducible methods and (2) a measure of accuracy.³⁹ For the first requirement, the methodology must be systematic and return the same result across different examiners. A high degree of methodological rigour is especially important for subjective methods because human judgment is vulnerable to numerous cognitive biases.⁴⁰ Examiners must follow a procedure that does not vary based on their subjective judgments. For instance, the fingerprint misidentification of the perpetrator of the 2004 Madrid train bombing case is widely considered an instance of a subjective methodology led astray by motivated reasoning.⁴¹ In that case, the FBI investigator reported a match with 100% certainty, but was

³⁸ Michael J Saks et al, "Forensic bitemark identification: weak foundations, exaggerated claims" (2016) 3 Journal of Law and the Biosciences 538-575.

³⁹ PCAST, *supra* note 37 at 47-54.

⁴⁰ Contextual Bias, *supra* note 37; NAS Report, *supra* note 6 at 122-124, 184-185; Goudge, *supra* note 30 at 387-390; PCAST Report, *ibid* at 31. For a review of the dangers posed by expert testimony, from a psychological perspective, see Michael J Saks & Barbara A Spellman, *The Psychological Foundations of Evidence Law* (New York: New York University Press, 2016) at 202-241.

⁴¹ PCAST, *supra* note 37 at 28; Robert B Stacey, "Report on the erroneous fingerprint individualization in the Madrid train bombing base" (2005) 7:1 Forensic Science Communications.

mistaken. Subsequent investigations found the investigator took advantage of an overly flexible methodology to come to the outcome he was seeking.

The DREP, while providing a useful systematic procedure, is still subjective. Akin to fingerprint identification, there is a procedure to follow. Each step, however, engages human judgment. Several steps are particularly subjective. For instance, in the horizontal gaze nystagmus test (HGNT),⁴² the suspect is asked to follow a stimulus (e.g., the officer's penlight), with the DRE looking for signs of "a lack of smooth pursuit".⁴³ That judgment – what is smooth and what is not – admits a great deal of subjectivity. Indeed, research has demonstrated that several contextual factors affect the reliability of the HGNT.⁴⁴ Similar subjectivity is found in the DREs' assessment of balance and the suspect's ability to divide his or her attention.⁴⁵

Moreover, a DRE's ultimate determination is a discretionary balancing of the totality of the steps.⁴⁶ As admitted by Constable Jellinek at Bingley's trial, there is no standardized scoring system or mechanism to weight the steps.⁴⁷ When examiners have such latitude, there is a greater chance that they will focus on the criteria that confirm their suspicions.⁴⁸ In fact, even highly-trained scientists have admitted to preferentially relying on and reporting the data points that

⁴² *Regulations*, *supra* note 8 at 2(a) and 3(b)(i).

⁴³ US Department of Transportation, Transport Safety Institute, National Highway Traffic Safety Administration, "Drug Evaluation and Classification Training" (January 2011 Edition) online: < http://www.duidcolorado.net/wp-content/uploads/2014/06/DRE_Manual_2011.pdf > at 94 [DRE Manual].

⁴⁴ For instance, time of day because the eye-fatigue may contribute to the ability to control one's gaze, see: Charles R Honts & Susan L Amato-Henderson, "Horizontal Gaze Nystagmus Test: The State of the Science in 1995" (1995) 71 ND L Rev 671-700 at 694-695 [Gaze Test].

⁴⁵ *Regulations*, *supra* note 9 at 3(c).

⁴⁶ DRE Manual, *supra* note 43 at 85.

⁴⁷ Bingley's Factum, *supra* note 10 at paras 11(g-h).

⁴⁸ Gary Edmond et al, "Thinking Forensics: Cognitive Science for forensic practitioners" (2016) *Sci Justice* at 3-4; Itiel E Dror, David Charlton & Alisa E Péron, "Contextual information renders experts vulnerable to making erroneous identifications" (2006) 156:1 *Forensic Science International* 74-78.

support their hypothesis when they have the latitude to do so (i.e., “researcher degrees of freedom”).⁴⁹

The accuracy (2) component of foundational validity can be measured in multiple ways. In criminal law, which seeks to avoid wrongful convictions,⁵⁰ the false positive rate (FPR),⁵¹ or the chance of declaring a hit when there was no hit, is especially important.⁵² In the context of drug testing, the FPR is the probability of detecting a drug when there was actually no meaningful amount present in the suspect’s system. The PCAST Report notes: “Methods with a high FPR are scientifically unreliable for making important judgments in court...To be considered reliable, the FPR should certainly be less than 5 percent and it may be appropriate that it be considerably lower, depending on the intended application.”⁵³

Alternatively, consider a method’s “sensitivity”,⁵⁴ which is the chance of reporting a hit when there is a hit in reality; in the DREP context, it is the probability of identifying a person under the influence of drugs when they are, in fact, under the influence of drugs. Sensitivity is a very important metric for investigative tools. For instance, imagine if DREs only identified 450 of 900 cases of actual inebriation, a sensitivity of 50%. In such cases, half of those drivers who were actually under the influence would evade the bodily fluid test and go free. Still another

⁴⁹ Leslie K John, George Loewenstein & Drazen Prelec, “Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling” (2012) 23:5 Psychological Science 524-532.

⁵⁰ Contextual Approach, *supra* note 4; *R v Hart*, 2014 SCC 52 at para 8, [2014] 2 SCR 544.

⁵¹ In other sciences, the FPR is known as “Type I Error”, see: Jacob Cohen, “A Power Primer” 112:1 Psychological Bulletin 155 -159 at 156 [Power Primer].

⁵² PCAST Report, *supra* note 37 at 151-152.

⁵³ *Ibid.*

⁵⁴ In other sciences, sensitivity is known as “statistical power”, see: Power Primer, *supra* note 51.

measure is the overall accuracy rate: the number of true positives and true negatives divided by the total number of cases.⁵⁵

It is important to note that these measures of accuracy are quantitatively distinct – a sensitive and overall accurate measure may still carry a high FPR. To make this point more concrete, consider this hypothetical example, the results of which are summarized in Table 1. Assume that in a population of 1,000 people tested, 900 have actually taken drugs and 100 have not. Now, assume that DREs identify 820 as having taken the drug and 180 as not. On these assumed facts, the DREs are demonstrating an overall accuracy of 88% and sensitivity of 88.89%. As shown in greater detail in Table 1, this is because overall accuracy is the number of true positives (800) and true negatives (80) divided by the total number of observations (1,000). And sensitivity is the number of true positives (800) divided by the total number of drug-influenced individuals (900).

An overall accuracy of 88% and sensitivity of 88.89% are impressive, but these metrics do not tell the whole story. On these assumed numbers, the same DREs also have a FPR of 20%. Recall that FPR is the number of false positives (20) divided by the number of people not actually under the influence (100). And in the criminal law context, a FPR of 20% is a cause for concern.

[Table 1 about here]

After determining that a methodology is foundationally valid, validity-as-applied must be established. In other words, it must be verified that the person performing the method in the

⁵⁵ Douglas J Beirness, Erin Beasley & Jacques Lecavalier, “The Accuracy of Evaluations by Drug Recognition Experts in Canada” (2009) 42:1 Canadian Society of Forensic Science Journal 75-79 at 78 [Beirness 2009].

current case is capable of accurately doing so in circumstances resembling that case. Subjective methods like the DREP are particularly at risk for failures of applied validity because of what psychologists call “individual differences”: humans vary on several dimensions related to drug recognition, like natural ability and motivation to study and work harder.⁵⁶ Measuring applied validity is typically accomplished through “proficiency testing”: the assessment of individual analysts to see if they can accurately employ the method.

Proficiency tests must themselves be performed rigorously.⁵⁷ The PCAST Report recommends that proficiency testing should be administered by arm’s length parties,⁵⁸ noting substantial deficits in proficiency testing in the fields of latent fingerprint analysis and firearms analysis (i.e., determining if ammunition is associated with a certain firearm).

Assuming that the DREP is foundationally valid, it still may be that some police officers are not good at applying it.⁵⁹ They may overweigh the results of certain parts of the test, or rely too much on a preconceived notion of whether the subject was drug-impaired. The latter, which is also known as confirmation bias – forcing incoming information to fit a pre-existing theory⁶⁰ – could flow from a variety of sources, such as the type of neighbourhood the suspect was detained in, the suspect’s race, or any other idiosyncratic contextual cue.⁶¹ Moreover, like any skill, the

⁵⁶ Bradford T Ulery et al, “Accuracy and reliability of forensic latent fingerprint decisions” (2011) 108:9 *Proceedings of the National Academy of Sciences* 7733-7738; Jonathan J Koehler “Fingerprint error rates and proficiency tests: what they are and why they matter” (2009) 59 *Hastings LJ* 1077; David White et al, “Perceptual expertise in forensic facial image comparison” (2015) 282:1814 *Proceedings of the Royal Society B*.

⁵⁷ PCAST Report, *supra* note 37 at 58.

⁵⁸ *Ibid* at 57.

⁵⁹ Although they did not use the term “proficiency test”, the importance of such tests accords with the Canadian Civil Liberties Association’s Intervener’s Factum at the Supreme Court of Canada, court file number 36610, available online; <http://www.scc-csc.ca/WebDocuments-DocumentsWeb/36610/FM030_Intervener_Canadian-Civil-Liberties-Asssocation.pdf> at para 27: “What if, in the officer’s experience evaluating individuals who are suspected of being impaired by drugs, that officer’s assessment is regularly not borne out the toxicological tests?”.

⁶⁰ Raymond S Nickerson, “Confirmation Bias: A Ubiquitous Phenomenon in Many Guises” (1998) 2:2 *Rev Gen Psychol* 175 at 175.

⁶¹ Contextual Bias, *supra* note 37.

ability to employ the DREP may erode over time. For example, the DRE may see his or her opinion corroborated by a subsequent urine analysis several times in a row and rely on this limited sample to become overconfident and thus not follow the method as rigorously as before. Or similarly, the DRE may begin to rely only on a subset of the 12 DREP steps.

Part III. Scientifically-Informed Legal Analysis

The above review demonstrates that, from a scientific standpoint, the DREP is a fundamentally subjective methodology. As a result, it raises a different set of concerns than more objective tests, such as breathalysers. Therefore, we contend that it was a mistake for *Bingley*'s Majority to place these types of evidence on such similar footing for two general reasons. First, due to its subjectivity, the DREP's applied validity (i.e., are the DREs accurately and reliably applying the method) is much more tenuous and thus requires more judicial oversight. And second, while the DREP may be good at detecting drug use (i.e., high sensitivity), its FPR appears to be considerably worse. This suggests that it is a useful investigative method, but a dangerous form of inculpatory evidence in court.

As an objective method, the results of a breathalyser demand a different type of scrutiny by courts. Like DNA analysis, breathalyser data is analyzed in a predominantly automated way, with little opportunity for subjective judgment.⁶² As a result, once foundational validity is established – perhaps through Parliament's review of empirical studies – the primary adjudicative issue is whether the methodology was properly followed. For example, was there a failure in the chain of custody?⁶³ Indeed, courts may exclude breathalyser results if the device

⁶² PCAST Report, *supra* note 37 at 46-47.

⁶³ *Ibid* at 50-51.

was not properly calibrated.⁶⁴ By analogy, DNA analysis of single-source samples are widely considered foundationally valid,⁶⁵ but can be unreliable if the methodology is not properly followed such as via “sample mix-ups, contamination, incorrect interpretation and errors in reporting.”⁶⁶ For instance, in March of 2017 the Attorney-General for Western Australia reported that 27 criminal convictions were in doubt for failure of a police lab to follow proper DNA testing procedures.⁶⁷

Subjective methods, which rely on human judgment, are qualitatively different. Whereas computer programs can be expected to run identically across crime labs and countries, human judgment varies a great deal by individual. Therefore, because Parliament deemed breathalysers “conclusive proof” and did not do the same with the DREP, it stands to reason that this was not mere inadvertence but a deliberate choice to treat different types of science differently. In other words, empowering the police to use the DREP should not be seen as endorsement of its unlimited use in court. The DREP is a much more fragile form of evidence.

The majority did not seem to recognize the importance of applied validity. In particular, they made a lack of applied validity a matter of weight, as opposed to admissibility: failing to “conduct the drug recognition evaluation in accordance with his or her training” is not a grounds for excluding the DREP evidence.⁶⁸ As peak scientific bodies have noted, applied validity is just as essential to the truth of an expert’s claim as foundational validity; the results of a

⁶⁴ *R v Carles*, 61 WCB (2d) 443 at paras 53-80, 2003 CarswellOnt 4962.

⁶⁵ PCAST Report, *supra* note 37 at 71-73.

⁶⁶ *Ibid* at 73.

⁶⁷ Graeme Powell, “Leading DNA scientist sacked, 27 criminal convictions in doubt, WA Attorney-General says” *ABC News* (2017) online: < <http://www.abc.net.au/news/2017-03-31/sacking-of-was-leading-dna-scientist-27-criminal-cases-in-doubt/8403618?pfmredirect=sm>>.

⁶⁸ *Bingley*, *supra* note 5 at para 32. Rather, the Majority held that the DRE’s evidence could only be excluded when its probative value was outweighed by its prejudicial effect. They suggested this may occur when a DRE was completely unable to explain how he or she came to the conclusion.

foundationally valid procedure mean nothing if the expert cannot or will not faithfully apply it. Moreover, as a matter of law, failure or inability to follow a methodology is a matter of admissibility, not just weight.⁶⁹

Finally, the absence of a known FPR associated with the DREP suggests that trial judges ought not abdicate their authority to assess its foundational validity and reliability. The Crown, in its submissions to the Court, noted that Parliament appeared to accept a study finding the DREP possessed a 98.6% accuracy rate.⁷⁰ A review of the parliamentary committee hearing transcript leaves it unclear whether that is an estimate of overall accuracy or sensitivity.⁷¹ But in any case, the committee's focus on this metric and lack of any evidence that they turned their mind to the FPR further supports a continuing gatekeeper role for trial judges in the context of DREs. This analysis coincides with the Dissent's broader point: Parliament seems to have only vetted the DREP's reliability as an investigative method,⁷² for which sensitivity and overall accuracy are the salient measures. And they did not consider whether it should provide confirmatory evidence to be presented directly to judges and juries, for which FPR is the most relevant measure.

Furthermore, and drawing again on Table 1, the state of the DREP science is, in fact, largely consistent with that table. Several studies have shown that DREs identify drug impairment with generally high overall accuracy and sensitivity.⁷³ But, at the same time, they do

⁶⁹ *R v Abbey*, 2009 ONCA 624 at para 119, 97 OR (3d) 330 [*Abbey*]; *R v Sriskanda*, 2016, 2016 ONCJ 667, CarswellOnt 18574.

⁷⁰ *Bingley*, *supra* note 4 at para 56.

⁷¹ Parliament of Canada, Standing Committee on Justice and Human Rights, 39th Parl, 2nd Sess, No 077 (14 June 2007) available online <
<http://www.parl.gc.ca/HousePublications/Publication.aspx?Language=e&Mode=1&Parl=39&Ses=1&DocId=3043489>>.

⁷² *Bingley*, *supra* note 5 at 46.

⁷³ *Beirness* 2007, *supra* note 13.

so with a false positive rate that ranges from 4%,⁷⁴ to 20%,⁷⁵ to as high as 57%.⁷⁶ One review of the existing research suggests that FPR rates are typically higher in studies when officers have reason to believe “that most participants had ingested drugs.”⁷⁷ The same can be said for the method as currently employed because DREs will only be called upon when the investigating officer has a reasonable suspicion of drug impairment. In fact, some Canadian trial courts have rejected the DRE evidence because of its high false positive rate.⁷⁸ A statutory interpretation exercise that was sensitive to the science of the DRE might have taken a closer look at whether Parliament considered the method’s FPR, and concluded Parliament did not intend such a fallible (but sensitive) method to evade judicial scrutiny.

Part IV. What happened to Contested Science?

Implicit in the above analysis was a normative argument: not only did the Majority fail to account for several scientific distinctions in its interpretive exercise, but in doing so they may have allowed unreliable science into trial courts across the country. This move counters a decades long trend that has increased the scope of scientific evidence that trial judges are directed to scrutinize: “The unmistakable overall trend of the jurisprudence, however, has been to tighten the admissibility requirements and to enhance the judge's gatekeeping role.”⁷⁹ This trend has moved in tandem with and taken notice of meta-scientific insights from works like

⁷⁴ George E Bigelow et al, “Identifying Types of Drug Intoxication: Laboratory Evaluation of the Subject Examination Procedure”, DOT HS 806 753, National Highway Traffic Safety Administration, Washington, DC, available online <http://decip.us/pdfs/Bigelow_1985_DRE_validation_study.pdf>.

⁷⁵ Beirness 2009, *supra* note 55 at 78.

⁷⁶ David Shinar & Edna Schechtman, “Drug Identification Performance on the Basis of Observable Signs and Symptoms” 37 Accident Analysis and Prevention 843–85.

⁷⁷ Beirness 2007, *supra* note 13 at 372.

⁷⁸ *R v Wood*, 2007 ABQB 503, 426 AR 335 [*Wood*]; *R v Steeves*, 2010 NBPC 25, 362 NBR (2d) 48; *R v Thomas*, 2012 BCPC 215, [2012] BCWLD 6042.

⁷⁹ *White Burgess*, *supra* note 1 at para 20.

Justice Goudge's Inquiry and the NAS Report. We will now briefly review this trend in law and suggest the Majority's decision in *Bingley* represents a step backwards.

The move towards enhanced gatekeeping of scientific evidence can be traced back to *R v Mohan*,⁸⁰ which still stands as Canada's leading expert evidence decision. Most are familiar with *Mohan*'s four well-trod requirements for expert evidence.⁸¹ But Justice Sopinka, writing for the Court in *Mohan*, also started Canadian jurisprudence down a new path for scientific evidence by holding that novel science should be subjected to "special scrutiny" as to its reliability.⁸² A year earlier, the U.S. Supreme Court had made the same move. It rejected deference to the general acceptance of science among experts, and instead demanded that trial judges assess the reliability and validity of the science prior to admitting it.⁸³ Justice Sopinka's decision, at least implicitly, aligned Canadian jurisprudence with that case: *Daubert v Merrell Dow Pharmaceuticals Inc.*⁸⁴

Any doubts about whether Canada had adopted *Daubert* were put to rest by the Supreme Court in *R v J (J-L)*.⁸⁵ In that case, the majority of the Court employed *Daubert*'s four factors as special scrutiny to exclude novel scientific evidence that purported to identify pedophilia on the basis of a psychological and physiological testing procedure.⁸⁶ Those factors are (1) whether and how the science has been tested; (2) whether and how the science has been peer-reviewed and

⁸⁰ *Mohan*, *supra* note 2.

⁸¹ They are: logical and legal relevance, necessity in assisting the trier of fact, absence of an exclusionary rule, and a properly qualified expert

⁸² *Mohan*, *supra* note 2 at para 32.

⁸³ For a concise review of *Daubert* and the U.S cases that preceded and followed it, see: David L Faigman et al, "Group to Individual (G2i) Inference in Scientific Expert Testimony" (2014) 81:2 University of Chicago Law Review 417-480 at 427-431.

⁸⁴ *Daubert*, *supra* note 32.

⁸⁵ *JLJ*, *supra* note 32.

⁸⁶ *Ibid* at para 33.

published; (3) the error rate associated with the science, and (4) general acceptance of the science in the field in which it is situated.⁸⁷

R v Trochym expanded *Daubert*'s ambit in Canada, noting that given science's evolutionary nature, even once-accepted theories and practices (i.e., non-novel) may be unreliable.⁸⁸

Therefore, even if it has received judicial recognition in the past, a technique or science whose underlying assumptions are challenged should not be admitted in evidence without first confirming the validity of those assumptions.

The majority then applied the *Daubert* factors to the science of hypnotically-recovered memories and ruled the memories derived from such procedures inadmissible, despite their use in previous decisions and the theoretical plausibility of the method that produces them.⁸⁹ Put simply, modern psychological science had found the method unreliable in application. In *White Burgess*, the Supreme Court reaffirmed the applicability of *Daubert* "in the case of an opinion based on novel or **contested science** or science used for a novel purpose".⁹⁰

The Majority's view in *Bingley* seems to undercut several decades of high court guidance directing trial judges to closely scrutinize scientific evidence. In particular, the Chief Justice stated that the DREP represented novel science,⁹¹ but exempted it from judicial scrutiny due to its place in the *Criminal Code*. This is troubling because the DREP fits more clearly into the category of *contested science*.⁹² Despite the original DREP studies dating back to 1985, several

⁸⁷ *Daubert*, *supra* note 32 at 593-595.

⁸⁸ *Trochym*, *supra* note 33 at para 32.

⁸⁹ *Ibid* at paras 55-61.

⁹⁰ *White Burgess*, *supra* note 1 at para 23 [**emphasis added**].

⁹¹ *Bingley*, *supra* note 5 at para 23.

⁹² *Trochym*, *supra* note 33; *White Burgess*, *supra* note 1 at para 23.

recent studies have demonstrated that the DREP suffers from a high FPR.⁹³ Furthermore, trial courts in Alberta, New Brunswick, and Nova Scotia have engaged in thoughtful examinations of the DREP and found it lacks a sufficient scientific foundation.⁹⁴ In particular, the Alberta Court of Queen’s Bench, in excluding DREP evidence, stated:⁹⁵

The low accuracy reported in the 1996 study is significant. Whether the cause is the DRE protocol’s inability to enable evaluators to accurately identify the physiological reactions caused by the particular class of drugs, the reliability of the science on which it is based, the evaluators’ training, or something else is well beyond the scope of this appeal. What is apparent, however, is the fallibility of the DRE, which strikes at the heart of its reliability.

But perhaps more troubling than the failure to characterize the DREP as contested science is the fact that the Majority seemed to altogether remove the “contested science” language from its enunciation of the type of science that will trigger *Daubert* scrutiny: “Such knowledge is **required only where the science is novel.**”⁹⁶ And while the Dissent did recognize the above case law as “unsettled”⁹⁷ and would have subjected the DREP to a *voir dire*, it also failed to pick up on the notion of contested science from *Trochym* and *White Burgess*.

This is not the time to narrow the scope of judicial gatekeeping. Recent meta-scientific insights from the NAS and PCAST have demonstrated that many longstanding assumptions about the scientific validity of the forensic sciences were simply wrong. And more generally, new findings from the basic sciences demonstrate that several fields have not adequately policed themselves and thus many established findings may be spurious.⁹⁸ Even old science can be

⁹³ Beirness 2007, *supra* note 13.

⁹⁴ See cases at note 78.

⁹⁵ *Wood, ibid* at para 82.

⁹⁶ *Bingley, supra* note 5 at para 22 [**emphasis added**].

⁹⁷ *Ibid* at para 57

⁹⁸ See sources at note 7.

invalid. This is all compounded by the fact that the “novel science” distinction proved unworkable. Some courts defined it as science not generally accepted by scientists,⁹⁹ while others thought it was science not yet accepted by courts.¹⁰⁰ Others defined it as temporally new science.¹⁰¹ The United States has abandoned the distinction for a contested science standard.¹⁰²

Part V. A (Scientific) Path Forward

Although never expressly stated, the Majority’s pullback on the trial judge’s gatekeeper role seems to have been driven by concerns about judicial economy. This consideration played a key role in the Crown’s submissions, which worried that exposing DREs to *voir dire* would promote “litigation in trial courts across the country”.¹⁰³ And, in fact, the procedural history in *Bingley* itself was long and included two trials.¹⁰⁴

In what was likely a response to these concerns about judicial economy, the Majority stated that before determining admissibility of expert evidence, the trial judge “must determine the nature and scope of the proposed expert opinion”.¹⁰⁵ As there was no issue of scope in *Bingley*, Chief Justice McLachlin’s statements suggest that carefully scoping expert evidence represents one of the courts preferred solutions to proliferating expert evidence. In fact, this approach traces back to Justice Doherty’s guidance in *Abbey*¹⁰⁶ and has now been twice highlighted by the Supreme Court.¹⁰⁷ We agree that this is a step in the right direction and

⁹⁹ *Wolfen v Shaw* (1998), 43 BCLR (3d) 190 at para 17, 76 ACWS (3d) 756.

¹⁰⁰ *R v Pearce*, 2014 MBCA 70 at para 69, 318 CCC (3d) 372.

¹⁰¹ *P(SF) v MacDonald*, 1998 ABQB 855 at para 14, 238 AR 175.

¹⁰² David Paciocco, “Context, Culture and the Law of Expert Evidence” (2001) 24 Advocates’ Quarterly 42-63 at 46.

¹⁰³ Respondent’s Factum at the Supreme Court of Canada, file number 36610, available online: <http://www.scc-csc.ca/WebDocuments-DocumentsWeb/36610/FM020_Respondent_Her-Majesty-the-Queen.pdf> at para 40.

¹⁰⁴ See *supra* note 18.

¹⁰⁵ *Bingley*, *supra* note 5 at paras 16-17, 29.

¹⁰⁶ *Abbey*, *supra* note 69 at paras 62-70.

¹⁰⁷ The other case was *Sekhon*, *supra* note 3 at paras 46-48.

believe it can be modified and applied to DRE evidence in a way that retains the trial judge's gatekeeper function.¹⁰⁸

We propose that, assuming the DREP must be considered foundationally valid pursuant to the Majority decision in *Bingley*, DREs could be questioned on a limited and predetermined set of questions focused on *validity-as-applied*.¹⁰⁹ In fact, the DREP presents what may be an ideal context for trial judges to perform such a function because DRE evidence is controlled by specific legislative provisions. In other words, it is much more predictable than typical expert evidence. These questions could be put to the DRE either at the preliminary stage of delineating the scope and nature of the expert evidence or as part of pretrial case management. If no *voir dire* is held in relation to this type of evidence, the questions could even form part of the trial.

A standard set of these applied validity questions for DREs might include the following: When were you last tested on your DREP proficiency by an independent third party? What were your results on this test? Did you follow all the required steps in applying the DREP? Do you tend to rely more on a particular aspect or aspects of the test? Were there any conditions that were not ideal for applying the procedure? Were you aware of any impediments to visibility? Were you aware of any extraneous noises or distractions?¹¹⁰

This suggestion represents an economical method that is sensitive to the science of the DREP. Perhaps more importantly, it does not abandon decades of jurisprudence focused on strengthening the trial judges' gatekeeping power. In the DREP context, it also prevents

¹⁰⁸ See Helena Likwornik, "Overstepping and Sidestepping: The Expert Evidence Dance" (2017) 35:4 Advocates' Journal 24-26, available online: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2840203>, which suggests that a requirement for the court to frame key questions for experts in advance of the trial may be useful in many contexts.

¹⁰⁹ Indeed, the Majority left open the possibility that DREs could still be excluded under the trial judge's final discretionary weighing of the costs and benefits of the evidence, see *Bingley*, *supra* note 5 at para 30.

¹¹⁰ Gaze Test, *supra* note 44.

professional groups (e.g., the IACP) from dictating what is reliable enough for Canadian courts. Of course, active involvement by the court continues the trend away from a strictly adversarial process. But this is a trend already underway through the gatekeeping process itself. The need for the trial judge to act as gate keeper, both initially and throughout the process, acknowledges that the adversarial process alone does not always offer sufficient protection against misleading scientific evidence.

Well before *Bingley* was decided, Chief Justice McLachlin provided the following quote in the introduction to the National Judicial Institute's *Science Manual for Canadian Judges*: "Without the proper tools, the justice system is vulnerable to unreliable expert scientific evidence."¹¹¹ While the Chief Justice was discussing scientific education when she wrote this introduction, the willingness and confidence to scrutinize investigatory methods is also an important tool. Prior to *Bingley*, trial courts around Canada were building a DREP jurisprudence, carefully considering whether it was reliable enough to serve as inculpatory evidence in criminal proceedings. This is a culture of judicial skepticism that should be fostered rather than tamped down, especially in this time of uncertainty within the forensic sciences.

¹¹¹ The National Judicial Institute, *Science Manual for Canadian Judges* (Ottawa: NJI, 2013) at 14.

Table 1

Table 1. In this hypothetical study of DRE accuracy, sensitivity is the number of true positives (800) divided by the total number of drug-influenced individuals (900) tested and thus 88.89%. Overall accuracy is the number of true positives (800) and true negatives (80) divided by the total number of observations (1000) and thus 88%. The FPR is the number of false positives divided the number of people not under the influence of a drug (100) and thus 20%.

		<u>DRE's Opinion</u>	
		<u>Positive</u> (820 total)	<u>Negative</u> (180 total)
Actual presence of drug	<u>Positive</u> (900 total)	800 (true positive)	100 (false negative)
	<u>Negative</u> (100 total)	20 (false positive)	80 (true negative)