An open and reproducible analysis of a controversial criminal law decision: Mission *IMM*-possible?

Jason M Chin

School of Law, University of Sydney, Australia

Institute for Globally Distributed Open Research and Education (IGDORE), Indonesia

ORCID: 0000-0002-6573-2670

I am indebted to Alex Holcombe, Simine Vazire, and Kathy Zeiler for the discussions that inspired and informed this article – and for their generous, collaborative spirit. I also thank Gary Edmond, David Hamer, and Andrew Roberts, my collaborators on the qualitative arm of this project.

Abstract

Reproducibility and open access are central to the research process, enabling researchers to verify and build upon each other's work, and allowing the public to rely on that work. These ideals are perhaps even more important in legal and criminological research, fields that actively seek to inform law and policy. This article has two goals. First, it seeks to advance legal and criminological research methods by serving as an example of a reproducible and open analysis of a controversial criminal evidence decision. Towards that end, this study relies on open source software, and includes an app (https://openlaw.shinyapps.io/imm-app/) allowing readers to access and read through the judicial decisions being analysed. The second goal is to examine the effect of the 2016 High Court of Australia decision, *IMM v The Queen*, which appeared to limit safeguards against evidence known to contribute to wrongful convictions in Australia and abroad.

Keywords

Reproducibility, open science, evidence, wrongful convictions

Part I: Introduction

> Good empirical work adheres to the replication standard: another researcher should be able to understand, evaluate, build on, and reproduce the research without any additional information from the author. This rule does not actually require anyone to replicate the results of an article or book; it only requires that researchers provide information-in the article or book or in some other publicly available or accessible form-sufficient to replicate the results in principle. Unfortunately, the present state of legal scholarship nearly always fails this most basic of tests. (Epstein & King 2002 p. 38)

Analysis of judicial decisions, quantitative or qualitative, is a fundamental methodology for understanding the law and its effects (Korobkin 2002 p. 1038). However, these analyses are often irreproducible, casting doubt on the inferences that can be drawn from them (Epstein & King 2002 p. 38). In much the same way, a large portion of criminology research appears to be irreproducible and criminologists have reported difficulty obtaining the data and materials underlying others' work (Burt 2020, Pickett 2020). In this article, I rely on new research from the meta-research and open science movement (Munafò et al. 2017) to provide, what, according to my knowledge, is the first fully open and reproducible quantitative analysis of judicial decisions. My focus is *IMM v The Queen* (2016) ('*IMM'*), a controversial criminal evidence law decision (see critiques and expressions of confusion about *IMM* in Edmond 2017, Hamer 2017, Roberts 2017, Odgers & Lancaster 2016, *Langford v Tasmania* 2018) that weakened trial judges' ability to exclude evidence known to contribute to wrongful convictions in Australia and abroad (e.g., forensic science, eyewitness identification).

Accordingly, I seek to advance two general goals in this article. First, I aim to move forward legal research and analysis methodology by providing an example of reproducible legal analysis (with open materials that can be reused and adapted by others). This includes

preregistration (see Part III), use of a PRISMA flow chart (Figure 2 and see also Part III), and open data and analysis scripts written with open source software (R). I provide all of the cases I rely on in an openly available app, the code for which is also open and can be reused by others (https://openlaw.shinyapps.io/imm-app/). The second aim is to provide a better understanding of a controversial and ambiguously worded decision, which seemed to have serious criminal justice consequences.

In what follows, Part II provides some background into *IMM* and the controversy surrounding it. Part III explains the need for reproducible legal analysis, the current lack of it, and the present study's methods. Then Part IV provides the results of my analysis, suggesting that from 2016-2020, *IMM* was read expansively and liberalized the admission of several types of dangerous evidence. Part V concludes with the study's limitations and future directions for reproducible legal research.

Part II. Safeguards against unfair and unsafe verdicts, *IMM*, and its controversy

In Australia (e.g., *R v Christie* 2014, *NSW Evidence Act* s. 137), as in other adversarial jurisdictions (in Canada see *R v Mohan* 1994, in the U.S. see *Federal Rules of Evidence* r. 403), trial judges can exclude evidence when its probative value (i.e., its ability to prove some fact in issue) is exceeded by its unfair prejudice. These provisions provide a safeguard against evidence that may be misused by the factfinder and produce unsafe verdicts. Unfair prejudice encompasses the possibility that the jury will misuse the evidence by assigning it more weight than it deserves. For example, courts in Canada and Australia (although this is now harder or potentially impossible in Australia, as we will see) sometimes exclude eyewitness identifications when they seem unreliable (*R v Holmes* 2002, *Dupas v R* 2012). This is done because honest but

mistaken eyewitness identifications are susceptible to being misused by a jury, a notion reinforced by the fact that they have contributed to wrongful convictions (in Australia see Dioso-Villa 2015 p. 182-183, in Canada see Denov & Campbell 2005).

Given the importance of this safeguard (excluding evidence when its unfair prejudice exceeds its probative value), the rules about how judges assess probative value are important. In *IMM*, the High Court of Australia (HCA), by a 4-3 majority, resolved two conflicting approaches held by Australia's two most populous states (New South Wales and Victoria) in a way that weakened the safeguard, and diverged from other similar jurisdictions (e.g., Canada and the U.S.). Prior to *IMM*, Victorian courts had construed probative value as including the evidence's reliability. In other words, evidence that was potentially unreliable could be excluded for low probative value because, according to Victorian courts, probative value included reliability. Including reliability in the probative value calculus matters because a lack of reliability is a major reason that evidence like eyewitness identifications and forensic science are dangerous (e.g., eyewitnesses are often mistaken, some forensic science practices are more error-prone than they seem).

In New South Wales (NSW), on the other hand, courts took a more hands-off approach by assuming the reliability of evidence when assessing probative value (i.e., taking reliability 'at its highest', *R v Shamouil* 2006 [50]) and trusting the jury to assign the appropriate weight to evidence. One issue with the NSW approach is that many reliability issues are not explored on cross-examination because the balance of resources often favours the prosecution (Findley 2008).

The *IMM* majority adopted the NSW approach, holding that both reliability and credibility should be taken at its highest when assessing probative value. This decision – albeit unclear, as we will see – is binding on *Australian Uniform Evidence Law* (‘*UEL’*) jurisdictions.[1]

The majority’s decision to take reliability at its highest attracted significant criticism. One branch of this criticism concerns apparent incoherence in the decision (giving way to potential difficulties in applying it, see Roberts 2017 p. 66-69). Notably, the majority seemed to articulate some difficult-to-understand exceptions to taking reliability at its highest. For example, it said that some evidence will be ‘simply unconvincing’ and thus lack probative value (*IMM v The Queen* 2016 [50]). Rather than explain what that meant, the majority provided an example of an eyewitness identification made on a foggy night, saying it had low probative value, not because it was unreliable, but because it was simply unconvincing. The problem with this example, one pointed out by at least one subsequent appellate decision (*Langford v Tasmania* 2018 [56]) and several legal researchers (Edmond 2017, Hamer 2017, Roberts 2017), was that the factors used by the majority to say this identification was simply unconvincing (e.g., poor visibility) are factors that go to reliability. This makes it difficult to understand what ‘simply unconvincing’ means and how subsequent courts should interpret it. The majority also stated that some evidence would lack probative value because it is ‘inherently incredible, fanciful or preposterous’ (*IMM* 2016 [58]), adjectives that seem to also evoke ideas of reliability.

Specific to tendency evidence (i.e., other similar acts of the accused that may suggest they committed the alleged act), the majority also held that such evidence could lack probative value if it was an allegation from the complainant about uncharged acts made by the accused

[1] *UEL* jurisdictions are: the Australian Capital Territory, Federal courts, Family courts, the Northern Territory, New South Wales, Tasmania, and Victoria. Queensland, Western Australia, and South Australia retain the common law.

against them (*IMM* 2016 [63]). *IMM* did little to explain why this evidence could lack probative value.[2] A subsequent decision acknowledged the confusion this part of holding caused and seemed to limit it substantially (*R v Bauer* 2018 [47]).

The other branch of criticism is that *IMM* seemed to open the door to evidence known to contribute to wrongful convictions. I have already mentioned the case with eyewitness identification. Gary Edmond (2017) and Justice Chris Maxwell (2019), have also raised the potential dangerous impact of *IMM* on the admissibility of forensic science evidence (Maxwell 2019 p. 642-643):

> In 2016, however, the High Court in *IMM v The Queen* held that the assessment of probative value […] did not involve any consideration of the reliability of the proposed evidence. The judge was required to assess probative value on the assumption that the jury would accept the evidence. […] On the present state of the law, therefore, the judge in a criminal trial is unable to perform the "gatekeeper" role as defined in [previous authorities]. The purpose of this article is to underline the critical importance of that role, and to highlight the urgent need for legislative intervention to reinstate it.

As Justice Maxwell notes, after *IMM*, there appears to be no scope for judicial exclusion of unsound forensic scientific evidence (unless it can be called 'simply unconvincing' or 'fanciful', a possibility assessed in Part IV), and therefore this area appears to call out for a legislated means to pave over the hole made by *IMM*. This contrasts with other jurisdictions that

[2] 'As is apparent from comparison of the trial judge's ruling with the Court of Appeal's reasons for judgment, **previous decisions of this Court have left unclear** when and if a complainant's evidence of uncharged sexual and other acts is admissible as tendency evidence in proof of charged sexual offences […] **It is unsatisfactory that trial judges and intermediate courts of appeal should be faced with that problem**.' (*R v Bauer* 2018 [47]) [**emphasis added**]

expose forensic science evidence to a reliability threshold (in Canada see Chin, Mellor & Growns 2019, in the U.S. see Mnookin 2018).

Part III. Open and reproducible legal analysis

> The field of metascience — the scientific study of science itself — is flourishing and has generated substantial empirical evidence for the existence and prevalence of threats to efficiency in knowledge accumulation. (Munafò 2017 p. 1)

Given the criticisms of *IMM* in Part II – along with its significance for the admissibility of a variety of types of prosecution evidence – I examined how *IMM* has been relied on in subsequent decisions. For example, how frequently are courts citing the credibility and reliability 'at its highest' ruling, and is it being applied beyond the context of *IMM* (which dealt with tendency evidence and evidence of a prior complaint made by the complainant)? How frequently have the apparent qualifications to that ruling (e.g., 'simply unconvincing') been considered? Have the parties relying on them done so successfully? Do jurisdictional differences remain? What types of potentially unreliable evidence is *IMM* being used to save, and do they fall into the categories that others have raised concerns about (e.g., expert evidence and forensic science)?

To address these questions, I conducted an open (i.e., the database and cases are publicly available) and reproducible (i.e., my analytic choices were thoroughly detailed and are also publicly available) quantitative analysis of cases citing *IMM* between the day it was handed down in April 2016 to when I began data collection in February 2020. Prior to describing my methods more fully, I will now briefly discuss the need for reproducible legal analysis in law, and especially criminal law. In particular, reproducible analysis improves the validity of the inferences made by the researcher. This is because reproducible research documents how and why decisions were chosen for analysis and why others were excluded so that users can

understand how well the observations support the inferences. Documenting this procedure also allows other researchers to verify the work and build upon it. Moreover, such analysis is more useful to practitioners because it provides confidence that the author has not omitted cases that could surprise them in court or before.

The need for open and reproducible legal analysis

The inferences drawn from analyses of judicial decisions have been called into question since at least 2002, when Epstein and King reported the results of studying 231 law journal articles from the prior decade (for a more recent critique, raising the same issues, see Zeiler 2016). Epstein and King (2002 p. 38) uncovered many methodological flaws in articles they reviewed, but perhaps most central was a lack of reproducibility. Reproducibility refers to reporting the article's methodology such that another could perform the same data collection (e.g., read the same statutes or decisions) and, without any further input from the authors, come to the same conclusion (Epstein & King 2002 p. 38; Hardwicke et al. 2020 p. 16). Epstein and King concluded (2002 p. 38) that, 'the present state of legal scholarship nearly always fails this most basic of tests'.

This conclusion held for both quantitative and qualitative analyses of judicial decisions because both seek to draw inferences from some subset of decisions. Epstein and King's criticism is worth reproducing in full because I will describe how this article's methods, using modern tools from the meta-research movement that were described in the epigraph that began this section, address their concerns (2002 p. 41):

> …the rule that empirical research must be replicable applies with equal force to studies relying on nonnumerical evidence. In many, perhaps most, instances, legal academics conducting these sorts of investigations rarely provide even a tracing of how they collected the evidence.

> Sklansky's essay on new originalism and Black's on state regulation of political parties are exemplary, but there are scores of other doctrinal studies that are equally negligent in providing the reader with guidelines sufficient to replicate the analysis. We rarely learn
>
> (1) How authors canvassed the relevant case law and what precisely was the population from which they sampled;
>
> (2) How authors selected their cases and how many they read;
>
> (3) How authors distinguished "key" or "a few ... exemplary cases" from those that are not central or not typical.

Since Epstein and King's article, a wave of research has emerged over the past decade aimed at making the scientific literature more reliable and accessible (this movement has sometimes been described as one focused on 'reproducibility' or 'open science'; it involves the study of research itself, or 'meta-research' (see Munafò et al. 2017; Allen & Mehler 2019; Vazire & Holcombe 2020; Chin, Mellor & Growns 2019). This work has produced new technology for hosting data and materials so that researchers are no longer confined by word limits in journals (National Academies 2018 p. 114-115). It has also innovated techniques and templates for describing how data was selected and excluded (i.e., Preferred Reporting Items for Systematic Reviews and Meta-Analyses or PRISMA, Moher et al. 2009). Although many of these new developments are applicable to legal research, they have almost never been applied to it (but see Chin, Lutsky & Dror 2018).

Methodology and methodological contributions

In the course of detailing the current article's methodology, I will also describe how I adapted three practices from the reproducibility movement. Those methodological advances are

(a) preregistration, (b) reporting excluded cases using a PRISMA diagram, and (c) open data, methods, and a judicial decision exploring app.

Preregistered scope and coding scheme

One risk when conducting research is that the researcher will exert some influence on the results in a way that was not disclosed (Munafò et al., 2017). I encountered this issue with an analysis of judicial decisions in which my colleagues and I found that a widely-discussed Supreme Court of Canada decision inspired a great deal of evidence challenges, but that the exclusion rate for the type of evidence in question barely changed after the target case (Chin, Lutsky & Dror 2019, Figure 1). This was an exciting finding, but we found that there were many ways we could have influenced our results through the cases we selected (e.g., shifting the time frame) or through how we defined evidence exclusions (e.g., sometimes in a bench trial, a judge would say they assigned evidence no weight, is that tantamount to an exclusion?). It is tempting to make these decisions in a way that makes the results seem more exciting (see also Epstein & King, 2002, 103-8).

[**Figure 1 about here**]

Preregistration helps combat researcher bias. Preregistration is public commitment to a data collection and analysis strategy before data collection begins (Nosek et al. 2018). It helps dissuade and make salient data-contingent changes to methodology because such changes will be easier to spot. For instance, determining data are outliers after observing their effect on the results can bias findings (Simmons, Nelson & Simonsohn 2011), but this is more difficult to do when rules for excluding outliers are preregistered. For clinical medical trials, preregistration is required by law in some jurisdictions because of the dangers of biased reporting in this field that

impacts public health (Dickersin & Chalmers 2011). Researchers in economics, psychology, and political science increasingly preregister their studies (Christensen et al. 2019). And, social science journals increasingly encourage or require preregistration (Center for Open Science 2020). Still, there is little evidence of preregistration among legal researchers (see Chin et al. 2020).

Prior to reading most of the cases in this study, I preregistered its scope and coding scheme (https://osf.io/uq7fk). As disclosed in the preregistration, I read the first 39 cases citing *IMM* prior to preregistering. I did this to determine whether it was feasible to discern from these cases whether evidence with reliability and credibility issues was being admitted in reliance on *IMM*, and whether I would have time to do this. Nosek and colleagues (2018 p. 2603) recommend disclosing the data that has already been collected prior to preregistration when it is not feasible to preregister prior to this.

Ultimately, as is preregistered and is publicly available (https://osf.io/uq7fk), I read all decisions citing *IMM* on the *Lexis Advance* Australian database, up to those published on February 13, 2020 (which was when I began reading and coding). This is a total of 244 cases. I was only interested in a subset of these decisions and so many did not receive any further analysis. In particular, my preregistered scope was limited to *UEL* decisions because I wanted to study *IMM*'s effect on courts that were, in theory, bound by the decision. Still, all decisions are included in the open database, so researchers interested in *IMM*'s broader effect can review those cases (https://osf.io/chvb9/). Moreover, and also as preregistered, I only further considered decisions that cited *IMM* for its rulings regarding how judges should assess probative value.

Of the 244 cases, 177 cases met my requirements for further analysis (see Figure 2 and below under 'Prisma Diagram'). Those cases contain 268 evidence decisions. The main aspects of these decisions I coded were (full codebook is available, https://osf.io/eg5m8/):

- general reference information (e.g., date of decision, neutral citation);
- why *IMM* was cited (e.g., reliability and credibility at its highest, the simply unconvincing qualification, the fanciful evidence qualification, the evidence of uncharged acts qualification);[3]
- the type of evidence under consideration (e.g., tendency, expert opinion);
- aspects about the evidence the court mentioned that could affect its probative value (e.g., delay between the witness seeing the event and giving evidence;
- whether the evidence was admitted or excluded;
- any reservations expressed about following *IMM*.

I performed all the coding, discussing difficult cases with a colleague and collaborator on a different arm of this project (David Hamer). I made one change to the coding scheme after coding began (and recoded prior cases to match it). Specifically, I changed the way I coded the variable mentioned above about factors affecting the evidence's probative value. I originally intended to code the reasons the court found the evidence had high or low probative value, but it was too difficult to do this because courts rarely expressly said which factors they were relying on. Rather than rely on my judgment, I changed the coding scheme (see preregistration

[3] As can be seen in the codebook, I provided a narrative description of why *IMM* was cited. These were then broken down into categories: assess credibility and reliability at its highest; the meaning of probative value; the meaning of significant probative value, *IMM*'s discussion of concoction's effect on probative value, the uncharged evidence from the complainant 'qualification', the simply unconvincing 'qualification', the fanciful 'qualification', and whether the evidence being subject to competing inferences reduced probative value.

amendment, https://osf.io/h6fyt/) to simply list factors mentioned in the judgment (whether or not the court expressly relied on those reasons) when assessing evidence's probative value.

PRISMA diagram

The second methodological advance I used and adapted for legal research was the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram (Moher et al. 2009).

The PRISMA flow diagram is one half of the PRISMA Statement (the other half being a reporting checklist, Moher et al. 2009). The PRISMA Statement was developed to address bias and lack of transparency in systematic reviews (e.g., a review of many studies to determine if a certain medical intervention is effective). Systematic reviews involve collection and analysis of data that already exists. As a result, there is a risk that, consciously or unconsciously, systematic reviewers will select studies that fit with the narrative they wish to support. In this way, they are akin to analyses of judicial decisions because judicial decisions are also pre-existing data at risk of being selected in a way that biases the results. As a result, the PRISMA flow diagram is of use to those doing doctrinal research. In fact, it gives structure to one of Epstein & King's (2002 pp. 103-6) primary recommendations to combat the irreproducibility they found 'Record the Process by Which Data Come to Be Observed'.

Figure 2 shows this study's PRISMA flow diagram. The initial search returned 244 cases. 18 were screened out immediately for being out of jurisdiction. A further 49 were screened because they did not involve assessment of probative value (67 were excluded out in total). 176 cases remained and those 176 contained 267 decisions evidence decisions. The reasons for all excluded cases can be found in the supplementary materials (https://osf.io/4zph9/). One example

of a non-probative value citation is *IMM* being cited help interpret the *Migration Act* (*Han v Minister for Home Affairs* 2018) because both the *UEL* and that act use the word 'significant'.

Open data and methods

Finally, making an article's data and materials openly available in a public repository is on the rise in economics, psychology, and political science (Christensen et al 2018). The National Academies of Science (2018 p. 108) recently described these practices as 'essential' to the modern movement towards open research. Indeed, these practices further the scientific and academic ideals of researchers verifying existing findings and building upon them (Munafò et al., 2017). Legal and criminological research appear to be lagging cognate fields in these areas (Zeiler 2016, Burt 2020), although I am not aware of systematic research confirming this (but see Chin et al. 2020).

The data for this study are available on a public repository (https://osf.io/chvb9/). This includes an app (https://openlaw.shinyapps.io/imm-app/) that allows easy browsing of the data and includes Austlii links to improve verifiability and usability by other researchers and practitioners. Moreover, the code used to produce the results in Part IV is also open (https://osf.io/jbfra/). This code also lays out all cases that were screened above in Figure 2. The code for the app is also open (https://osf.io/fwc6z/). These resources may be of use to researchers undertaking similar projects.

Part IV. Results

I will now review some general trends I observed in this sample of evidence decisions that rely on *IMM*. I will provide only descriptive statistics because I do not seek to make inferences beyond these cases. Moreover, as preregistered, I will focus on broad trends, and

refrain from analysing the facts and reasoning in specific cases. Rather, that is something I hope other researchers and practitioners will leverage this database to do.[4] Readers can also follow the R Markdown (i.e., analysis file) that documents how the below results were generated, using the same headings as below (https://osf.io/jbfra/).

Why was *IMM* cited? Credibility and reliability its highest, and qualifications to that rule

First, a look at the reasons *IMM* was cited suggests that its main legacy has been to weaken the safeguards against unreliable evidence unduly affecting trials (Table 1). More specifically, looking at the first row of Table 1 (all decisions), 77% cited *IMM*'s wording directing judges to take reliability and credibility at its highest. Less common were citations to its guidance about what probative value means (31%) and what significant probative value means (in the context of tendency evidence) (2%).

[**Table 1 about here**]

In contrast to that widespread citation of reliability and credibility at its highest, citations to the qualifications to that have been rare – and, even rarer have been decisions finding that qualification was made out. As can also be seen in the first row of Table 1, only about 11% of decisions cited the fanciful qualification and 14% cited the simply unconvincing qualification. Further, about 3% cited the passage suggesting evidence of uncharged acts from the complainant may have low probative value.[5] Note, however, this analysis cannot help explain whether parties are simply not raising these qualifications or, if they are, they are not making it into published

[4] I am involved in one such project examining evidence that judges concluded were 'simply unconvincing' and thus excludable.

[5] I am including this part of the judgment because it was a large source of confusion (see above) that later had to be clarified in *Bauer*. However, it does not fit cleanly into this analysis as a qualification to credibility and reliability at its highest because the majority was never clear about why evidence of uncharged acts from the complainant would have low probative value.

decisions. In any event, when these qualifications are raised, they rarely succeed. Evidence was found to be fanciful and incredible in less than 1% of cases and simply unconvincing in about 3%. Similarly, the uncharged acts from the complainant argument was accepted in 3% of decisions.

Another way to explore the role of *IMM* is by examining the percentage of times the qualifications were accepted when evidence was excluded (the 3rd line in Table 1). When broken down this way, we see that the rules that seem to allow some assessment of reliability (fanciful and simply unconvincing) only make up about 10% of the exclusions (evidence of an uncharged act from the complainant makes up another 10%). This suggests that while evidence is still being excluded (88 decisions excluded evidence in this sample), it is typically for other reasons than for the qualifications laid out by the *IMM* majority.

Jurisdictional differences

Recall that *IMM* resolved two conflicting approaches. Prior to *IMM*, the Victorian position allowed courts to find evidence lacked probative value because it was unreliable (*Dupas v R* 2012). NSW courts did not (*R v Shamouil* 2006). As a result, I investigated whether there is evidence that NSW has more enthusiastically applied the credibility and reliability at its highest ruling and less enthusiastically applied the qualifications (and alternatively, does a streak of judicial activism remain in Victoria?).

[**Figure 3 about here**]

The pattern of citing *IMM*'s evidence at its highest ruling – and not the parts dealing with the limits of that rule – appears stronger in NSW (Table 2). Indeed, NSW citations to the

evidence at its highest holding occurs in over 90% of its decisions, the highest of any jurisdiction.

[**Table 2 about here**]

A similar picture arises when examining whether the qualifications to evidence at its highest were made out (Table 2, Figure 3). This data, displayed in Figure 3, suggest that, in this sample, NSW courts were less likely to rely on the inroads that the *IMM* majority provided to excluding evidence; such decisions made up a smaller proportion of NSW's decisions than in other states.

Cases with credibility and reliability issues

It may also be informative to consider the decisions broken down by whether there was a credibility or reliability issue with the evidence mentioned somewhere in the judgment. These decisions are listed in Appendix A. The cases illuminate several reliability and credibility issues that have been linked to wrongful convictions. For instance, in admissions cases, *IMM* seemed to facilitate admitting evidence when the person making the admission was suffering from a mental illness (*R v Lou* 2017, *R v Fantakis* 2018). In one of the largest studies of false confessions in the U.S., 43% of exonerees who had made a false confession were mentally ill (Garrett 2010 p. 1064). With regard to expert evidence, *IMM* has been cited in cases in which experts were challenged because they were performing a subjective analysis but were exposed to biasing information (*Chen v R* 2018), and when they failed to disclose the basis of their forensic scientific opinion (*Langford v Tasmania* 2018; *R v Volpe* 2018).[6] Cognitive bias and lack of a

[6] Although note *Volpe* was reversed in a decision outside the timeframe of this study because, according to the appellate court, the trial court did not properly assess unfair prejudice, see *Volpe v R* 2020.

scientific basis are two common concerns with expert evidence (National Research Council 2009 p. 8, 122-125).

[**Table 3 about here**]

Dividing the decisions based on whether there was a credibility or reliability issue also helps validate the coding and analysis in this study (see Table 3). In other words, we see that when there was a credibility or reliability issue, the court was more likely to refer to *IMM*'s evidence at its highest ruling (about 90% when there was a credibility or reliability issue, and about 70% when there was not), and it was also more likely to consider the qualifications to that rule. This suggests the study's coding was accurate, picking up on the underlying mechanisms for why courts refer to *IMM* (and why a party may plead it).

Type of evidence

Next, consider the type of evidence *IMM* was applied to. Recall that *IMM* itself concerned the probative value of tendency and complaint evidence. Accordingly, it was possible that subsequent courts would be reluctant to impose *IMM*'s rulings on assessing probative value to other types of evidence, seeing the guidance as less well suited to other types. On the other hand, the majority took no steps to limit the judgment: 'While *IMM* was not concerned with scientific, medical or technical evidence, the decision appears [to apply broadly]' (Edmond 2017 p. 112).

Table 4 contains applications of *IMM*'s evidence at its highest ruling broken by the type of evidence being considered. This table confirms Edmond's concern. Tendency and complaint make up the plurality of *IMM* applications (about 30%), but it has been applied to a panoply of evidence. These non-tendency and complaint applications are relatively spread out across several

types of evidence that researchers and public inquiries (see Dioso-Villa 2015, Ontario Ministry of the Attorney General 1998) have flagged as dangerous in criminal matters (e.g., admissions, consciousness of guilt, expert opinions, eyewitness identifications).

From Table 4, we can also see that courts considering the admissibility of eyewitness identifications and general eyewitness evidence, are the most likely to have considered *IMM*'s 'simply unconvincing' qualification to evidence at its highest (~80% of eyewitness identification decisions and ~65% for other eyewitness decisions). This suggests the incoherence and indeterminacy in the simply unconvincing qualification has restricted subsequent decisions. Recall that the majority did not provide much explanation of what simply unconvincing meant, other than to provide the example of an unreliable eyewitness identification (while at the same time saying reliability had nothing to do with it, *IMM v The Queen* 2016 [50]). As a result, it seems that subsequent courts and litigants have focused on that analogy to an eyewitness, rather than apply the rule more broadly or extend it to other potentially 'unconvincing' evidence (e.g., an error-prone forensic technique).

Expressions of disapproval and other comments

A look at the decisions in which the court expressly stated some resistance or confusion in applying *IMM* may also be useful. Some of these questioned the logic of excluding tendency evidence when it was the complainant's evidence of uncharged acts (*Packard v R* 2018, *R v O'Brien* 2017). However, as noted, this was eventually clarified in *Bauer*.

The rest seemed to struggle with the seemingly vast implications of taking credibility and reliability at its highest when assessing probative value. In *BM v R* (2017), the court suggested that despite *IMM* not expressly saying so, the possibility of concoction and contamination

(seemingly a reliability and credibility issue) could no longer could reduce evidence's probative value. Similarly, in *R v Zarshoy* (2017), the court suggested that assigning lower probative value to evidence because it is open to multiple inferences may have to be rethought after *IMM* (ostensibly because evidence open to multiple inferences is less reliable). Finally, a Tasmanian appellate decision pointed to that apparent contradiction between taking reliability at its highest, yet discounting unreliable identifications (*Langford v Tasmania* 2017 [56]):

> An aspect of the reasons of the majority in *IMM* which is somewhat difficult to reconcile with the comments from those reasons quoted above, is their Honours' reference at [50] to the "weak identification" example and the suggestion that the deficiencies surrounding an identification would impact on the assessment of the probative value of such evidence. On one view, the aspects of an identification which make it weak are matters which affect the reliability of the evidence, and, hence, ought not be considered when assessing the probative value of the evidence…

Outside the *UEL*

Finally, it is worth noting that, while non-*Uniform Evidence Law* cases were not considered in any of the above analyses (Figure 2), *IMM* does seem to have been impactful outside of those jurisdictions it is binding on (see full dataset, https://osf.io/chvb9/). In other words, in Australian common law states, *IMM* does not restrict courts from considering evidence's reliability when assessing probative value. However, South Australian courts have cited *IMM* approvingly, and in some cases, seemingly authoritatively (*R v LM* 2018 [28], *R v PAR* 2016 [85]) when declining to take reliability into account. Western Australian courts have picked up on *IMM*'s guidance about lowering the probative value of evidence of uncharged acts from the complainant (*WA v SGG* 2017, *RMD v WA* 2017, *WA v CGT* 2017, *WA v TKR* 2017).

And, for reasons that are not immediately clear, *IMM* was not cited in Queensland for the duration of the present study.

Part V. Conclusion: Challenges in performing reproducible legal and criminological research

> The challenges to reproducible science are systemic and cultural, but that does not mean they cannot be met. (Munafò et al 2017 p. 7)

As Munafò and colleagues say in the above epigraph, reproducible research is not always easy. In crime research, for instance, current incentive structures and peer review practices do not reward reproducibility (Burt 2020). Within legal research, it appears normative to make claims about the state of the law or trends in judicial decisions without describing and explaining the basis of those claims (Epstein & King 2002). This is true of research that makes an empirical claim in a more casual way (e.g., saying 'courts tend' to decide a certain way and citing a few cases without explaining from what population those cases were chosen, how representative they are, and why others were excluded).[7] It is also true, however, of articles claiming to have performed more systematic research of judicial decisions, where the authors may provide examples of the keywords they searched for, without providing all of them or how they decided to exclude some cases from analysis. An example of this is saying '[d]atabases were searched using terms including...' (Edmond 2019 fn. 4, see also Farahany 2016).

Another challenge with reproducible legal research, especially when starting without established procedures and precedents to follow, is that it takes time. The present project was quite resource-demanding when it came to reading and studying decisions (although researching

[7] Epstein and King (2002 p. 98-99) provide several examples of these practices. See also Patterson (2009 p. 939): '…courts tend to look for procedural injustice before granting relief…'.

authorities is an important part of much legal research), coding them, writing the scripts to analyse them, and writing the scripts to display them. These time demands can prevent potentially usefully research from reaching consumers in a timely way. This is especially important in legal research where the law is constantly changing. Indeed, important developments related to the effect of *IMM* have already occurred outside of this study's scope (see *Hague v The Queen* 2019, *Volpe v The Queen* 2020). Still, the data provided in the current project can be useful in comparing newer trends to those that occurred during this project's scope.

I also hope the methods and open materials within will make future projects more efficient. They can also learn from some of this study's mistakes and inefficiencies. For instance, I used *Lexis Advance* to develop this project's dataset. I did this because *Lexis* provides an easy-to-save document of all cases citing another case that I used as a checklist. In hindsight, I should have taken the time to use Austlii's similar function because, as a free resource, that would be easier for others to use to reproduce this study.[8] Future projects should also record cases that were difficult to code and require consultation, or use two coders and measure agreements and disagreements. This would be easier on smaller-scale projects because it is already time consuming for one person to code cases.

Other developments from the meta-research movement could assist legal researchers, both generally and in tackling some of the difficulties mentioned above. Curated repositories of resources and training materials are available for researchers interested in making their work more accessible and reproducible (e.g., the Open Scholarship Knowledge Base, OER Commons

[8] Note, however, the project's app provides Austlii links.

2020 and the Framework for Open and Reproducible Research Training, FORTT 2020). Best practice guides may also assist in this respect (Klein et al. 2018, Meyer et al. 2018).

While the above limitations and challenges are serious, we should not lose sight of the fact that openness and reproducibility are essential to the validity and usability of legal and criminological research. These are fields of research that seek to influence law and policy. Their results, therefore, should be fully assailable and reusable by others. The current project, its data, materials, and even limitations, may provide a small step towards those ideals.

References

Allen C, Mehler DM (2019) Open science challenges, benefits and tips in early career and beyond. *PLoS biology* 17(5): e3000246.

Burt C (2020) Doing Better Science: Improving Review & Publication Protocols to Enhance the Quality of Criminological Evidence. *The Criminologist* 45(4): 1-6.

Chin J, DeHaven AC, Heycke T, Holcombe AO, Mellor DT, Pickett J, Steltenpohl CN, Vazire S, Zeiler K (2020) Improving the credibility of empirical legal research: practical suggestions for researchers, journals, and law schools. *LawArXiv*. https://doi.org/10.31228/osf.io/952gh.

Chin JM, Growns B, Mellor DT (2019) Improving expert evidence: the role of open science and transparency. *Ottawa Law Review* 50(2): 365-410.

Chin JM, Lutsky M, Dror IE (2019) The Biases of Experts: An Empirical Analysis of Expert Witness Challenges. *Manitoba Law Journal* 42(4): 21-67.

Christensen G, Wang Z, Paluck EL, Swanson N, Birke DJ, Miguel E, Littman R (2019) Open Science Practices are on the Rise: The State of Social Science (3S) Survey. *MetaArXiv*. https://doi.org/10.31222/osf.io/5rksu.

Denov MS, Campbell KM (2005) Criminal Injustice: Understanding the Causes, Effects, and Responses to Wrongful Conviction in Canada. *Journal of Contemporary Criminal Justice* 21(3): 224-249.

Dickersin K, Chalmers I (2011) Recognizing, investigating and dealing with incomplete and biased reporting of clinical research: from Francis Bacon to the WHO. *Journal of the Royal Society of Medicine* 104(12): 532-538.

Dioso-Villa R (2015) A Repository of Wrongful Convictions in Australia: First Steps toward Estimating Prevalence and Causal Contributing Factors. *Flinders Law Journal* 17(2): 163-202.

Edmond G (2017) Icarus and the Evidence Act: section 137, probative value and taking forensic science evidence at its highest. *Melbourne University Law Review* 41(1): 106-154.

Edmond G (2019) Latent science: A history of challenges to fingerprint evidence in Australia. *University of Queensland Law Journal* 38(2): 301-365.

Epstein L, King G (2002) The Rules of Inference. *The University of Chicago Law Review* 69(1): 1-133.

Farahany NA (2016) Neuroscience and behavioral genetics in US criminal law: an empirical analysis. *Journal of Law and the Biosciences* 2(3): 485-509.

FORTT (2019) Introducing a Framework for Open and Reproducible Research Training (FORRT). *OSF Preprints*. https://doi.org/10.31219/osf.io/bnh7p.

Garrett BL (2009) The substance of false confessions. *Stanford Law Review* 62(4): 1051-1119.

Hamer D (2017) The Unstable Province of Jury Fact-Finding: Evidence Exclusion, Probative Value and Judicial Restraint after *IMM v The Queen*. *Melbourne University Law Review* 41(2): 106-154.

Hardwicke TE, Serghiou S, Janiaud P, Danchev V, Crüwell S, Goodman SN, Ioannidis JP (2020) Calibrating the scientific ecosystem through meta-research. *Annual Review of Statistics and Its Application* 7: 11-37.

Klein O, Hardwicke TE, Aust F, Breuer J, Danielsson H, Mohr AH, IJzerman H, Nilsonne G, Vanpaemel W, Frank MC (2018) A practical guide for transparency in psychological science. *Collabra: Psychology* 4(1): 1-15.

Korobkin R (2002) Empirical Scholarship in Contract Law: Possibilities and pitfalls. *University of Illinois Law Review* 4: 689-726.

Maxwell C (2019) Preventing Miscarriages of Justice: The Reliability of Forensic Evidence and the Role of the Trial Judge as Gatekeeper. *Australian Law Journal* 93(8): 642-654.

Meyer MN (2018) Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science* 1(1): 131-144.

Moher D, Liberati A, Tetzlaff J, Altman DG, Prisma Group (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS med* 6(7): e1000097.

Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, du Sert NP, Simonsohn U, Wagenmakers E, Ware JW, Ioannidis JPA (2017) A manifesto for reproducible science. *Nature Human Behaviour* 1: 1-9.

National Academies of Sciences, Engineering, and Medicine (2018) *Open science by design: Realizing a vision for 21st century research*. Washington, DC: National Academies Press.

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

Nosek BA, Ebersole CR, DeHaven AC, Mellor DT (2018) The preregistration revolution. *Proceedings of the National Academy of Sciences* 115(11): 2600-2606.

Odgers S, Lancaster R (2016) The probative value of evidence. *Bar News* 36-43.

OER Commons (2020) Open Scholarship Knowledge Base. https://www.oercommons.org/hubs/oskb.

Ontario Ministry of the Attorney General (1998) *The Commission on Proceedings Involving Guy Paul Morin: Report*. Toronto: Queen's Printer for Ontario.

Paterson J (2009) The Australian unfair contract terms law: The rise of substantive unfairness as ground for review of standard form consumer contracts. *Melbourne University Law Review* 33(3): 934-956.

Picket JT (2020) The Stewart Retractions: A Quantitative and Qualitative Analysis. *Econ Journal Watch* 17(1): 152-190.

Roberts A (2017) Probative Value, Reliability, and Rationality. In Roberts A and Gans J (eds) *Critical Perspectives on the Uniform Evidence Law*: 62-79. Sydney: The Federation Press.

Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22(11): 1359-1366.

Vazire S, Holcombe AO (2020) Where Are The Self-Correcting Mechanisms In Science?. *PsyArXiv*. https://doi.org/10.31234/osf.io/kgqzt.

Zeiler K (2016) The Future of Empirical Legal Scholarship: Where Might We Go from Here. *Journal of Legal Education* 66(1): 78-99.

Case List

BM v R [2017] NSWCCA 253.

Chen v R [2018] NSWCCA 106.

Dupas v R [2012] VSCA 328.

Hague (Wearne) v R [2019] VSCA 218.

IMM v The Queen [2016] HCA 14.

Langford v Tasmania [2018] TASCCA 1.

Packard v R [2018] VSCA 45.

R v Bauer [2018] HCA 40.

R v Christie [1914] AC 545.

R v Fantakis [2018] NSWSC 1814.

R v LM [2018] SADC 92.

R v Lou [2017] ACTSC 127.

R v Mohan [1994] 2 SCR 9.

R v O'Brien [2017] NTSC 34.

R v PAR [2016] SADC 85.

R v Shamouil [2006] NSWCCA 112.

R v Volpe (1) [2018] VSC 796.

R v Zarshoy [2017] NSWSC 1437.

RMD v WA [2017] WASCA 70

Volpe v R [2020] VSCA 268.

WA v CGT [2017] WADC 55.

WA v SGG [2017] WADC 47.

WA v TKR [2017] WADC 66.

Legislation

NSW (New South Wales). 1995. *Evidence Act*, No 25.

US. *Federal Rules of Evidence*, 28 USC.

Figure 1. Example of a quantitative analysis of judicial decisions

The number of relevant decisions (i.e., expert challenges) before and after *White Burgess Langille Inman v Abbott and Haliburton Co* 2015 SCC 23 charted against the number exclusions per year (*White Burgess* was decided in 2015). This is the author version of a figure presented in Chin, Lutsky & Dror 2019).

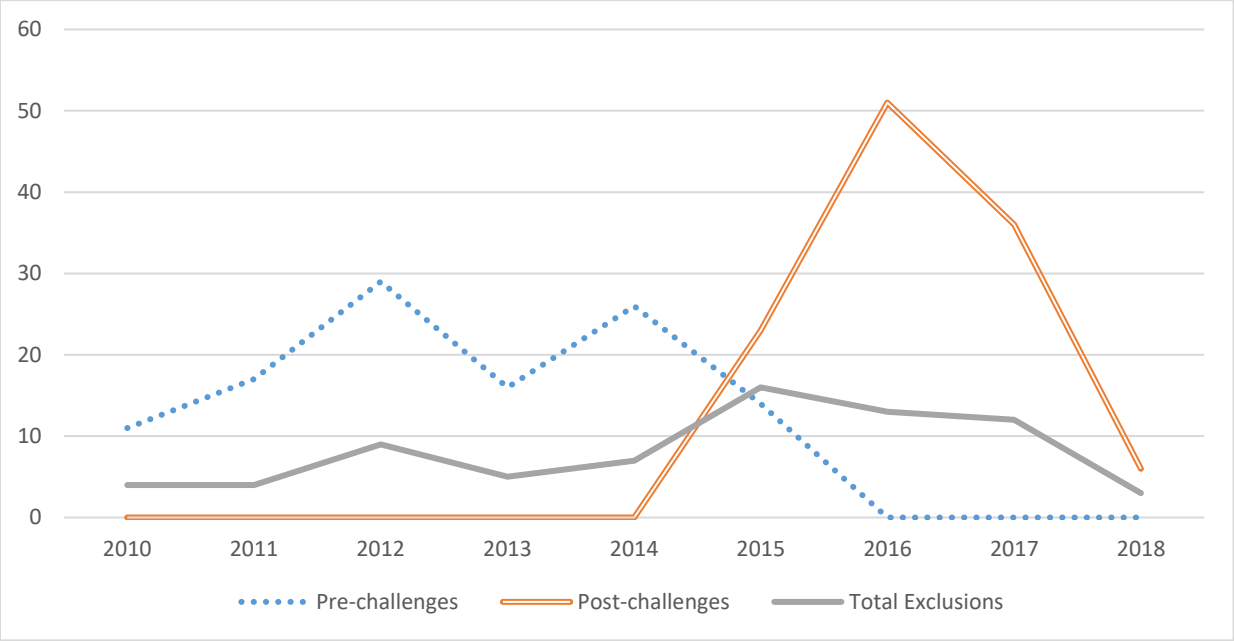

Figure 2. PRISMA Flow Diagram

Figure 2. The PRISMA flow diagram (Moher et al 2009) adapted for legal analysis. 18 cases were screened for being out of jurisdiction. A further 49 were excluded according to the preregistered criteria for not citing *IMM* about the assessment of probative value. Full reasons for exclusions are available in the supplementary material (https://osf.io/4zph9/).

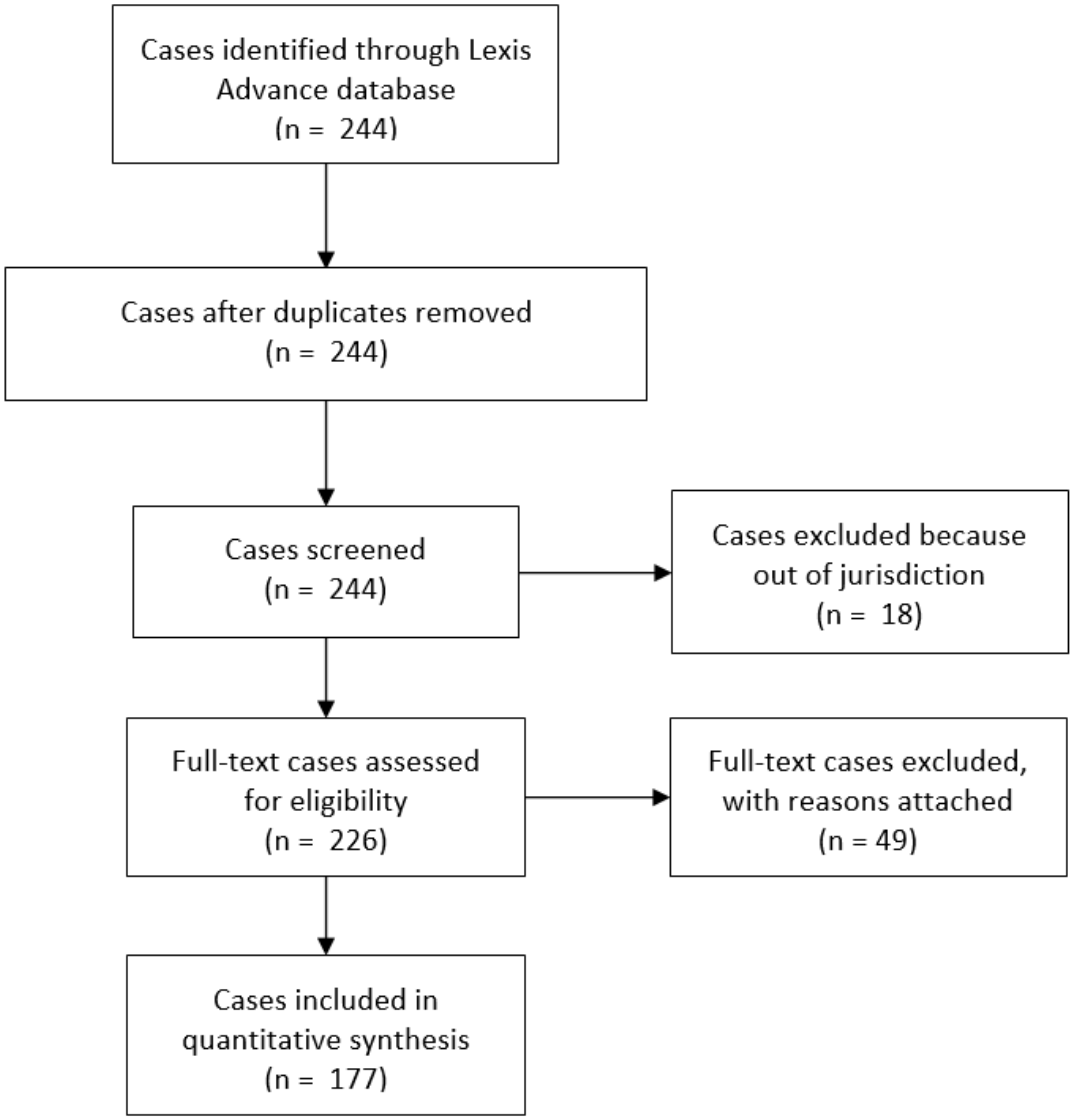

Figure 3. Citations to *IMM* by State

Figure 3 displays the number of decisions citing *IMM* for four reasons across seven states. First, are cases that cite *IMM* for its decision regarding taking reliability and credibility at its highest when assessing probative value. The other three are apparent qualifications to that rule: when evidence is fanciful, when it is simply unconvincing, and tendency evidence that comes from the complainant about uncharged incidents. Only citations in which the court accepted the qualification was made out on the facts are included in Figure 3.

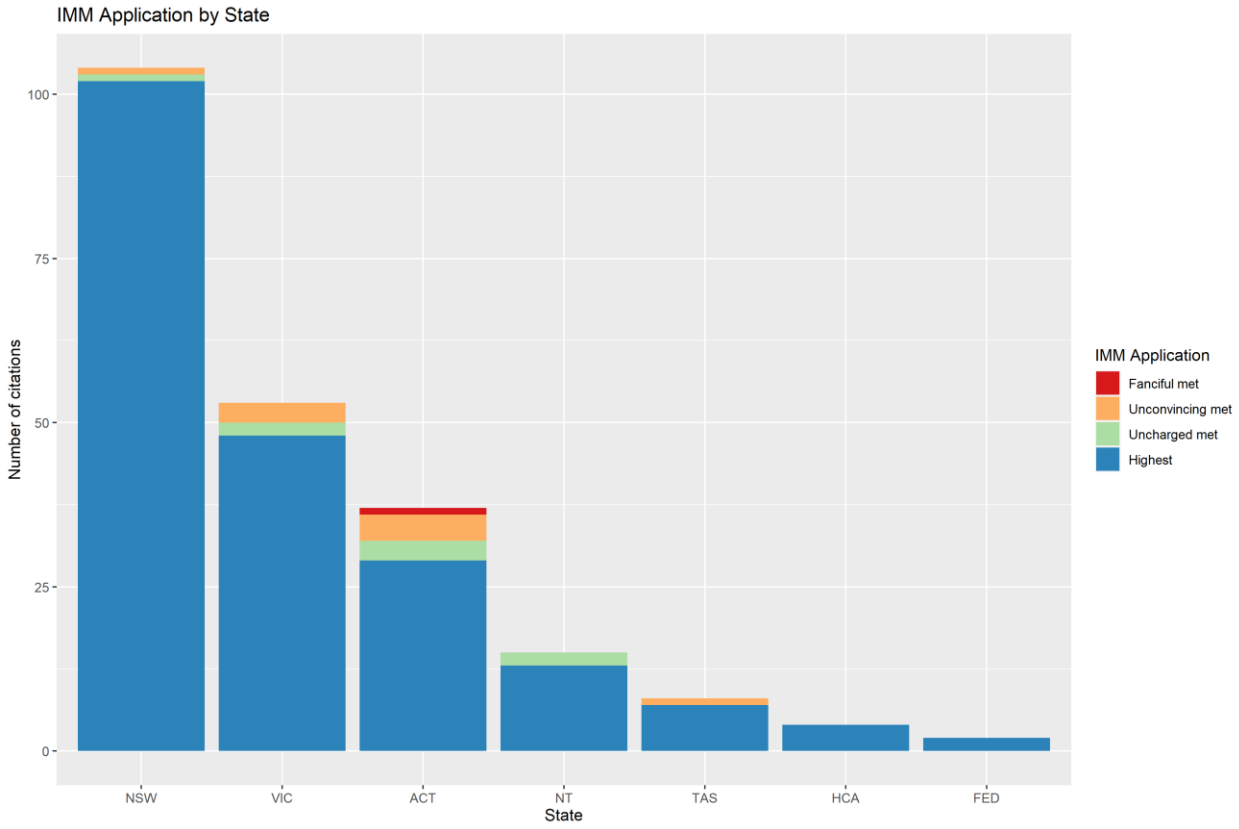

Table 1. Summary of evidence decisions by whether the evidence was admitted or excluded

Table 1 describes the way in which decisions in this sample cited *IMM*. The first row is all decisions, the second is just those in which the evidence was admitted and the third is those in which the evidence was excluded. The second column is the number of decisions. From there, Table 1 presents the percent of those decisions that cited *IMM* for: reliability and credibility at its highest, its explanation of significant probative value (for tendency and coincidence evidence), its definition of probative value, the fanciful qualification to evidence at its highest, the simply unconvincing qualification, the evidence of uncharged incidents from the complainant, and then whether those three qualifications were made out on the facts of the case.

Admitted?	# Decisions	% Highest	% Sig. PV	% Def. PV	% Fanciful	% Uncon	% Uncharged	% Is Fanciful	% Is Uncon	% Is Uncharged
All	268	76.5	31	1.5	10.8	13.8	7.8	0.4	3.4	3
Yes	180	79.4	25.6	2.2	10	12.8	6.1	0	0	0
No	88	70.5	42	0	12.5	15.9	11.4	1.1	10.2	9.1

Table 2. Summary of evidence decisions by State

Table 2 contains the same information as Table 1, broken down by state. It corresponds to Figure 3.

State	# dec	% Admit	% Highest	% Fanciful	% Uncon	% Uncharged	% Fanciful	% Is Uncon	% Is Uncharged
TOT	268	67.2	76.5	10.8	13.8	7.8	0.4	3.4	3
NSW	112	69.6	91.1	11.6	8	1.8	0	0.9	0.9
VIC	59	72.9	81.4	1.7	18.6	13.6	0	5.1	3.4
ACT	51	60.8	56.9	21.6	27.5	13.7	2	7.8	5.9
NT	28	57.1	46.4	10.7	0	14.3	0	0	7.1
TAS	8	62.5	87.5	12.5	25	0	0	12.5	0
FED	5	40	40	0	0	0	0	0	0
HCA	5	100	80	0	20	0	0	0	0

Table 3. Summary of evidence decisions with credibility or reliability issues

Table 3 displays evidence decisions, as divided by whether there was a credibility or reliability issue with the evidence mentioned in the decision. It presents: the number of decisions, the percent of those decisions that admitted the evidence, the percent that cited *IMM* for reliability and credibility at its highest, percent that cited the fanciful qualification, and the percent that cited the simply unconvincing qualification.

Credibility or reliability issue?	# decisions	% Admit	% Highest	% Fanciful	% Uncon
No	164	64.6	68.3	7.3	6.7
Yes	104	71.2	89.4	16.3	25

Table 4. Evidence decisions by type of evidence

Table 4 displays evidence decisions citing *IMM* broken down by the type of evidence under consideration. The table also provides the number of decisions for each type of evidence and percent that represents. The final column is the percent of decisions of that particular type in which the court considered the simply unconvincing qualification. Only types of evidence that appear five or more times in the dataset are included. Further details on how evidence type was decided can be found in the supplementary materials (https://osf.io/eg5m8/).

Type of evidence	# Decisions	% Decisions	% Simply unconvincing considered
admission	15	5.6	6.7
hearsay	15	5.6	13.3
eyewitness	14	5.2	64.3
general inculpatory	22	8.2	22.7
complaint	5	1.9	40
expert opinion	14	5.2	7.1
coincidence	9	3.4	11.1
consciousness of guilt	14	5.2	0
context/relationship	15	5.6	13.3
tendency	73	27.2	9.6
eyewitness identification	9	3.4	77.8

Appendix A. Cases with reliability or credibility issues admitted through *IMM*

Appendix A is available at https://osf.io/cnp7q/.