MEASURING JUSTICE[†]

William Cullerne Bown

The possibility of measuring the success of the criminal justice system in distinguishing the guilty from the innocent is often dismissed as impossible or at least impractical. Here I claim to demonstrate that such epistemic measurement would only be difficult. All measurement consists of two steps, the acquisition of observations and their processing through a computational framework. The law has lacked both, but I have recently put forward a computational framework and here I set out how the relevant observations can be obtained. This completes the conceptual foundations necessary for the development of jurisprudence as a social science, for policymaking in the law that is rooted in rational concern for epistemic outcomes, and for us to fulfil the modern, democratic promise that our forebears found in Blackstone's ratio.

We don't attempt to quantify the law's success or failure in convicting the guilty and acquitting the innocent. Perhaps the lack of interest in empirical measurement of the epistemic outcomes of the criminal justice system stems from the common belief that it is impossible. For example, Epps wrote recently, '... because of the very nature of the problem at issue, we have no effective way to measure the phenomenon empirically'.(Epps, 2014, p. 1145)

This article will argue that we can measure outcomes and, to convince, I will explicitly set out the methods to be used. That work is the primary purpose of this article and, for reasons of space, only very briefly at the end will I make the argument that the quantification of concerns that measurement allows would be a good thing. More useful perhaps at the outset is to remind ourselves of three immediate consequences of the lack of measurement.

First, within the academy, lack of measurement of what is a central concern is one of the things that separates jurisprudence from the social sciences.

Second, Porter has documented how it was the army not teachers that insisted on the quantification of IQ tests; tax collectors not companies that insisted on codification of accounting practices; and politicians not engineers who insisted on cost-benefit analyses. The role of quantification in such cases is to act as a 'technology of trust' in far-flung

[†] William Cullerne Bown, *Measuring Justice*, International Journal of Evidence and Proof. Copyright © 2019 SAGE Publications. Reprinted by permission of SAGE Publications, http://journals.sagepub.com/home/epj. Accepted for publication in April 2019. This is the version accepted for publication and can be downloaded from my personal webpage at https://quantitativejurisprudence.com.

I would like to thank Federico Picinali for inspiring feedback on an earlier draft of this article.

democracies. Without measurement, no similar oversight can be established over the law, denying it an important source of democratic accountability and legitimacy.(Porter, 1996)

Third, as Tribe noted, when we think of justice, the lack of mathematics shifts our focus from outcomes to ritual.(Tribe, 1971, p. 1393)

In short, lack of measurement is a fundamental fault line. If we could measure epistemic outcomes, we might expect the law to become a very different thing.

Given its reliance on numbers, the kind of measurement described here bears comparison to attempts to apply principles of probability or economics to the law. However, it is distinct from those two siblings and has its own underpinning theory.(See for example Hand, 2004) To clarify what it entails, when we talk of measuring something, we usually think of making *observations*. Yet these are meaningless if they are not embedded in a *computational framework* that allows them to be evaluated. Often this framework is so simple that we don't even notice it; as Hand has noted, measurement is often invisible.(Hand, 2004, p. 1) For example, suppose you are investing $10 and with god-like insight know that policy X would result in the first set of results in Figure 1 and policy Y the second set. Which outcome is better?

Figure 1: A simple policy choice

	$
X	100
Y	1000

Answer: Y. Furthermore, we have a general rule that we can apply to any two options: the bigger number is better.

Now let us turn to the law. Suppose you are trying 10158 cases and with god-like insight know that policy X would result in the first set of results in Figure 2 and policy Y the second set. Which is better? More importantly, what is the rule that would allow you to decide between *any* two sets of results?

Figure 2: A more difficult policy choice

	True Conviction	**False Acquittal**	**True Acquittal**	**False Conviction**
X	560	1300	7913	385
Y	288	1572	8197	101

Now the answer is far from obvious. Thus it becomes clear that when considering measurement the law has not one but two problems: a) the well-discussed one of

observations, of knowing which verdicts are true and false; and b) the hidden one of establishing a computational framework for evaluating such observations according to a linear scale of what is often called *effectiveness*.

Dividing the act of measurement into two parts clarifies the nature of the difference between the law and the natural sciences, for we can see that Mother Nature has turned her back on us not once but twice. She has declined to provide us with a computational framework, what Van Rijsbergen called an 'empirical ordering'; and she has declined to provide us with what we might think of as an 'empirical verdict' on our decisions that she offers to, for example, the makers of a pregnancy test kit. However, the law is not the only discipline that suffers from these two deficiencies. The same is true of Van Rijsbergen's field of information retrieval, the discipline of indexing and search engines, and it has acquired trustworthy measurements enabling spectacular progress over the past 50 years.(Van Rijsbergen, 1979 Chapter 7)

Just as Doll did not need to know *how* smoking caused lung cancer to establish that it *did*, we will not need to rely on any of the competing explanations of the internal workings of a court or the rest of the criminal justice system. The system will be treated here as a kind of black box, a *binary classifier*, akin to a pregnancy test or a software algorithm.(Doll & Hill, 1956)

Hand sees all measurement as lying on a spectrum between extremes of *representation* (so that a measure of length refers to a physical object such as a metal bar) and *pragmatism* (where the economic concept of gross domestic product cannot be divorced from its means of measurement). Most measurement in Hand's view involves aspects of both and from this point of view we can say that observations of the four tallies would be representational (as false convictions, for example, will be counted against the "absolute" scale of the integers themselves) while their combination through a computational framework into a single number would, thanks to the lack of an empirical ordering, have to involve a degree of pragmatism.(Hand, 2016)

Inspired by Laudan (who was inspired by recent progress in meta-ethics), we can also frame the problem in epistemological terms.(Laudan, 2006) The standard of proof can be taken as an example of a *rule* or policy governing the *epistemology* of the law, the means by which a court decides on a verdict. The basis on which the standard of proof is determined is then a *meta-rule* and part of the law's *meta-epistemology*. It is then an open question whether the law today has a meta-epistemology at all; to obtain one, it needs a coherent meta-rule that allows it to decide between any two policy options. Thus a computational framework for determining whether policy X or Y in Figure 2 is better is equivalent to a meta-rule and becomes a quantitative definition of the good.

I have recently put forward an example of such a computational framework which, while not unique, claims to be uniquely appropriate for the law. Given suitable observations and a value for b^2, the exchange rate at which we are prepared to trade false acquittals for false convictions, it allows us to always determine whether X or Y is the better policy. But the development of this meta-rule has left outstanding the other part of the measurement problem, the need to obtain suitable observations, which is the focus of this article.(Cullerne Bown, 2018)

What is required is a methodology of observation, but this cannot simply be described; the case has to be made for its appropriateness and practicality. I begin in Section 1 by arguing that the law makes measurements already, but in what I call *subjective* fashion. I argue that the assertion that measurement is impossible entails a kind of doublethink and that the question is not how to measure but how to measure better.

In Section 2 I briefly set out the new computational framework I have put forward and define the kinds of observation that it makes necessary. In Section 3 I tackle the problem of how to estimate one of these, the rate of wrongful convictions. In Section 4 I look at the qualities that would make such estimates trustworthy, following Kaye in breaking our interest in the quality of the jury system into three distinct concerns.(Kaye, 1980, p. 1013) To decide between policies we need observations at different levels of confidence and in Section 5 I examine different ways these could be acquired. In Section 6 I provide a worked example of a computation, partly based on existing data for rape in the UK. In the Conclusion I clarify and argue in favour of my emphasis on outcomes rather than procedure, partly by arguing that invisible principles of measurement are already deeply entrenched in the law, a case I make with the help of the sixteenth century woodcut in Figure 3.

Figure 3: Creating an instrument to measure length in feet
The jury-like procedure set out in this woodcut and text is taken from a surveyor's manual originally published in German in 1535 by Jacob Koebel, a mathematician and publisher based in Heidelberg.

'Stand at the door of a church on a Sunday and bid 16 men to stop, tall ones and small ones, as they happen to pass out when the service is finished; then make them put their left feet one behind the other, and the length thus obtained shall be a right and lawful rood to measure and survey the land with, and the 16th part of it shall be the right and lawful foot.' [1]

1. Subjective Measurement

Consider the following question: are most of those we convict truly guilty?

One answer, which would be true to Epps's scepticism, is, 'I have no idea.' But who would be willing to endorse that position? It amounts to saying that, when considering outcomes, we have no idea at all whether our system is just or unjust, that we have no epistemic basis for any policy. Though the procedural or ritualistic aspect of justice is understandably emphasised in jurisprudence this should not blind us to the fact that justice is generally

[1] Jacob Koebel, *Geometrei. Von künstlichem Feldmessen, vnd absehen Allerhandt Höhe, Fleche, Ebne, Weitte vund Breite*. The woodcut is taken from an edition of 1598 available in facsimile online at http://digital.slub-dresden.de/werkansicht/dlf/8074/14/; the translation of the text is taken from (HAND, 2016, P. 3)

considered to be firstly a question of outcomes. Is anyone really prepared to abandon *all* claims to accuracy?

A more common response I think would be, 'Yes, I believe so.' Then, how has that belief been acquired? Is it merely magical thinking with no basis in reality, so that – so far as outcomes go – policymaking in the law is no more than witchcraft in wigs? If not, then it must be ultimately rooted in some genuine knowledge of the outcomes of the criminal justice system, which entails some kind of measurement.

This last seems to me the most realistic account. However, the kind of measurement involved is evidently very different from using a thermometer to measure temperature. We can make a comparison with probability, where Franklin has given an account of its history that runs for more than a millennium before quantification (largely through the medium of legal reasoning). In the same way, we can allow for forms of measurement of legal outcomes that are non-quantitative, and may have been going on for a long time.(Franklin, 2002)

I call this subjective measurement and an example of it can be found in Dahlman's recent depiction of how policies governing the burden of proof are made:

> When reports indicate that too few guilty defendants are convicted, the criminal justice system reacts by making it easier for the prosecution to reach the required degree of probability. When reports indicate that too many innocent defendants are convicted, the criminal justice system reacts in the opposite direction, and makes it harder for the prosecution.(Dahlman, 2018, p. 15)

Dahlman's depiction illustrates what seem to me the four distinguishing features of subjective measurements: i) a reliance on *ad hoc* reports that may be first- or second-hand (or indeed more remote); but which lack both ii) a formal process of sampling and tallying; and iii) a quantitative methodology by which to translate the tallies (or other statistics) into evaluations of policy options; so that iv) we have no meta-rule and decisions on policy questions can only be made on an intrinsically subjective basis.

Features i) and ii) are concerned with observations and I refer to them as *informality*; feature iii) is concerned with how the observations are processed via some kind of computational framework to provide guidance on policy options and I refer to it as *incoherence*; feature iv) is the overall subjectivity. The informality in Dahlman's depiction is I think is clear on any reading. The incoherence of the methodology he outlines needs a little bit of work to establish.

Dahlman's "reports" do not have to be anchored to epistemic outcomes at all. The word may be interpreted as reflecting mere gossip or lobbying. Such an interpretation leads us back to witchcraft in wigs. But if the policy decisions that are ultimately made reflect some

genuine grasp of outcomes, then the reports themselves must have some kind of genuine relationship with outcomes, what we might call, using the word loosely, a correlation.

Next note that, as the procedure is founded on the assumption that the reporter knows which defendants are truly guilty, saying 'too few guilty defendants are convicted' is logically identical to the more common phrasing of 'too many guilty defendants are acquitted'. So then Dahlman's 'too few' and 'too many' indicate a methodology that boils down to the selection of a number that represents a ratio of (reports of) false convictions to false acquittals, that is a number akin to Blackstone's ratio. Above this number, shove policies in one direction, below this number, shove them the other way, which amounts to turning the number itself into an objective. This perhaps seems to be a meta-rule of the kind we are seeking but is doubly incoherent. First, despite its apparently enormous implications, nobody knows what this number is (or agrees what it should be). Second, even if we did know it, the uselessness as a meta-rule of a ratio of this kind when used as an *objective* has been established by many authors.(Cullerne Bown, 2019, p. 14; See for example these three: Laudan, 2006, p. 72; Risinger, 2008, p. 1003) One problem is that the ratio says nothing about the two true outcomes, and hence a single ratio is compatible with all kinds of systems, from anarchy to totalitarianism.

Thus Dahlman's depiction of how decisions are made is in my view half right, half wrong. It is right in that this kind of shoving – first one way, then the other – is how things are done. It is wrong in that this does not, as Dahlman seems to me to suggest, comfortingly imply that the direction of the shove is informed by some kind of rational, coherent methodology rooted in outcomes. There is a discomforting and fundamental incapacity in the law here, a conclusion which the reader may find easier to accept by the end of this paper once the scale of the machinery required to establish a coherent methodology is clear.

In consequence of the informality and incoherence, the ultimate decisions on policy that Dahlman refers to can only be made on a subjective basis. One consequence of this is that a transparent account of a policy decision that connects the decision with its epistemic grounds and consequences cannot be given, either to the public or to other policymakers. It is a kind of democratic deficit that reflects the law's lack of access to Porter's technology of trust.(Porter, 1996)

We can use this analytical framework to assess the significance of empirical work. For example, Risinger's use of DNA evidence to assess the accuracy of historic convictions for rape-murders in the United States is striking, his observations clearly formal rather than informal. However, even with such empirical data we still lack a meta-rule that would allow us to derive policy from them and so the problem of methodological coherence remains and any policy decision we make must remain intrinsically subjective. Thus a subtlety of our

current subjective methods is that reasoning may involve statistics that are, in themselves, well compiled but which are asked to perform a role in argument for which there is no methodological justification. (Risinger, 2006)

To return to the assertion of impossibility exemplified by Epps, we can see now that it involves a kind of doublethink. On the one hand, subjective – which is to say, deeply compromised – measurement goes on all the time. On the other hand, the possibility of better, objective measurement is dismissed out of hand.

Measurement is not, as is often supposed, impossible or impractical. On the contrary, it is a process of fundamental importance. Both legal scholars and policymakers draw conclusions about how the criminal justice system does and should work from measurement all the time, but such reasoning is not easily seen precisely because the measuring remains unacknowledged.

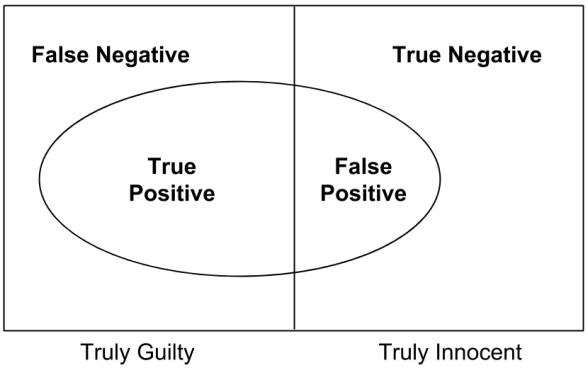
Just as metaphysics is meta to physics so measurement is meta to policymaking so far as it seeks an epistemic justification. Thus, to paraphrase Whitehead's dictum concerning scientists and metaphysics, every jurist in order to preserve her reputation has to say measurement is impractical; what she means is she dislikes having her measurements scrutinised.(Conger, 1927)

In short, the question is not how can we measure; it is how can we measure better.

2. The Computational Framework

To briefly summarise the approach I have previously set out, we can imagine counting up the number of truly guilty people convicted (the true positives or TP) and falsely acquitted (false negatives, FN), and the number of innocent people righly acquitted (true negatives, TN) and falsely convicted (false positives, FP). If we count the conviction of two people for the same crime as two elements, then each element in our universe consists of a person-action pair. These four tallies can then be represented in the characteristic Venn diagram shown in Figure 3 that can be called the *4-gram.*

Figure 4: The 4-gram
The inner oval of this traditional Venn diagram contains the convictions, the enclosing rectangle the universe of acts under consideration.



For all the 4-gram's apparent simplicity and the voluminous literature discussing its components, there is no agreement on the scope of the universe defined by the enclosing rectangle. For someone who adopts the one-trial, jury's view of things, such as Kaplan, the universe and the 4-gram do not exist. Otherwise, should this be the universe of cases brought to court (a position explicitly adopted by Tribe in rejecting Kaplan and often implicit elsewhere), the wider one of those dragged into the system by the police (the basis for Packer's statistics and explicitly articulated by Risinger recently) or the widest possible one of everything that happens in society as a whole (Laudan)? (Kaplan, 1968; Laudan, 2006; Packer, 1968; Risinger, 2008; Tribe, 1971)

If we are considering what is best for our jursidiction as a whole, then the answer is clear, it must be to consider the complete picture. Otherwise we are leaving the problem of, for example, rapes that are never reported to someone else to fathom. We are then in a position analogous to that of a manufacturer of a diagnostic test or owner of a software algorithm that classifies documents, as opposed, for example, to a user of the test or algorithm. Since our quantitative frame of reference does not need Hart's sophistication and can be perfectly well expressed in terms of the commands of Bentham's sovereign, I call this the *sovereign* perspective.(Hart, 1973)

To begin measuring, the first things we need are the rates of the two kinds of error the system can make, false positives and false negatives. The errors can be measured against the yardsticks of either the true positives or the true negatives and different disciplines such as medical diagnostics and information retrieval have chosen differently. It is a question of appropriateness. The case can then be made for the true positives on the grounds of achievement and measurability. Briefly:

> Achievement – if the criminal justice system convicts a murderer (a true positive) it has achieved something; if it does not convict someone who is not a murderer (a true negative), it has achieved nothing. *Given* that someone has fallen under suspicion,

then it certainly does achieve something. But that presumption is the starting point for a court and its smaller universe of cases brought to it, not a sovereign.

Measurability – the true negatives include actions where no crime is committed and no conviction is obtained and the police never hear about it, an uncountable miasma that includes innocuous handshakes.

This choice then yields two ratios known as *precision* and *recall* which vary between 0 and 1:

$$P = Precision = \frac{TP}{TP + FP}$$

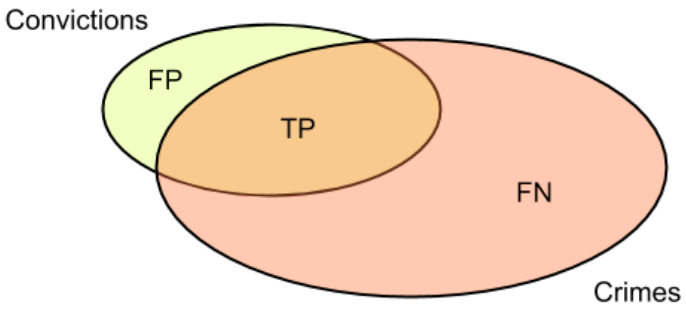
$$R = Recall = \frac{TP}{TP + FN}$$

Precision is the percentage of people we convict who are in fact guilty. Recall is the percentage of the guilty that we end up convicting. The two can be thought of as competing forms of security, from crime itself (the consequence of low recall) and false imprisonment (the consequence of low precision).

Since we have excluded the true negatives, our universe reduces to a countable *3-gram* as in Figure 5 composed only of true positives, false positives and false negatives.

Figure 5. The 3-gram

The most useful way of looking at this diagram is that we have a set of crimes and a set of convictions; their overlap, the true positives, is our achievement; and the bits that don't overlap are the two kinds of error.



To go beyond the previous work, in order to obtain estimates of precision and recall, there will be three steps:

- count the number of convictions: M = TP+FP
- estimate the number of crimes: N = TP+FN
- estimate the proportion of convictions that are true: P = Precision.

From these we can then derive recall:

$$R = Recall = \frac{M \times P}{N}$$

For the avoidance of doubt, since we have adopted the sovereign perspective, N here refers to the number of crimes in society as a whole.

We can then calculate the F_β-measure, a function of precision and recall:

$$F_\beta = \frac{(1+\beta^2)PR}{\beta^2 P + R}$$

where the value we assign to β is a way of making precision more or less important than recall.(Cullerne Bown, 2018)

The formula can also be re-written in terms of tallies as:

$$F_\beta = \frac{(1+\beta^2)TP}{(1+\beta^2)TP + \beta^2 FN + FP}$$

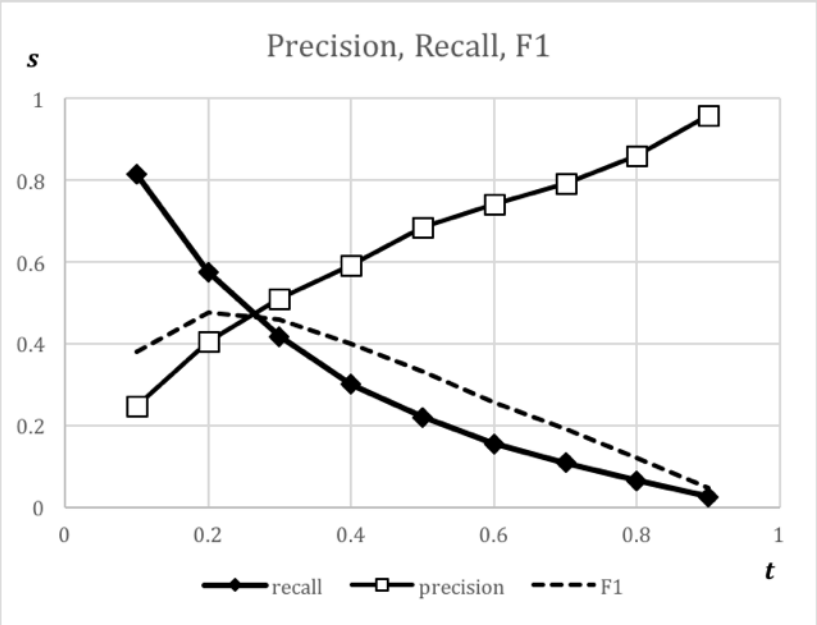
The false positives and false negatives here only appear in the denominator. Examining this shows that the number we assign to β^2 , b^2, functions as an exchange rate, so that 1 false negative and b^2 false positives have the same impact on the result and hence the same value to us. For example, if we choose the value $b^2 = 1/10$ then we have Blackstone's ratio and have decided that 1 false positive has the same value to us as 10 false negatives; and policies that lead to higher scores on the $F_{1/\sqrt{10}}$-measure are then considered better than policies that lead to lower scores.[2]

The F_β-measure for our chosen value of b^2 then is a *de jure* indicator of epistemic effectiveness. The choice of b^2 defines the good and the F_b-measure allows us to determine whether policy X or policy Y is better. The reasons for choosing the F_β-measure rather than another function were established originally in information retrieval. However, it clearly has a special attraction for the law as it can be seen as a methodology that finally allows Blackstone's ratio (or some other value judgement about the relative importance of the two kinds of error) to become properly consequential.(Van Rijsbergen, 1979)

[2] The ratio here is exactly the same as the one we uncovered in Dahlman's depiction; the difference is the use to which it is put, here an input (to a formula), there a desired output (of the system).

The curves for precision and recall and the F_β-measure when plotted against the confidence threshold, *t*, typically follow the general shape shown in Figure 6. As precision rises, recall falls, and vice-versa. The F_β-measure has a single maximum, although the value of *t* where that occurs will vary according to the choice of b^2. There is no analytic, mathematical proof that these features will always be present; they have however been empirically observed in a very wide range of binary classifiers in many different fields and, like those fields, we will take them as read.

Figure 6: Precision, recall and the F_1-measure
The chart plots precision, recall and the F_1-measure against a confidence threshold for a version of the Maui algorithm used to classify text documents. It shows that the F_b-measure with $b^2 = 1$ is maximised at the confidence threshold of 0.2. Thus if we value precision and recall equally, 0.2 is the best confidence threshold to use with this classifier.(Medelyan, 2009)



For this calculation of the F_β-measure to make sense, all the observations must refer to a single universe, which can be thought of as a single cohort defined by a period of time. So the tally of convictions, for example, must be for crimes *committed* during the same period as our estimate of the number of crimes.

In terms of data acquisition, M, the number of convictions, is to hand. N, the number of crimes, can only be obtained through surveying. As a challenge, this is of a kind with which we are already familiar. For example, the Crime Survey for England and Wales published by the Office for National Statistics seeks to establish the incidence of some crimes through surveys of households rather than from administrative data. Such crimes include crimes of

violence against children aged between 10 and 15. For the year ending March 2017, the CSEW offers estimates based on different two methodologies, each with a range based on a 95 per cent confidence interval. The *preferred* method indicates there were 359,000 incidents with a range of 292,000 to 426,000. The *broad* method indicates 660,000 incidents with a range of 570,000 to 750,000.('Crime in England and Wales - Office for National Statistics: year ending March 2017', 2017 Table UG10)

Another example of successful surveying, concerning rape, is provided below in the worked example.

From these examples we can conclude that the number of crimes is in principle measurable, at least in some cases. It is not clear that all crime is capable of being surveyed in this way, and certainly there are many types of crime for which surveys are currently considered unsuitable. The Crime Survey of England and Wales does not address drugs crimes, so-called vice crimes, financial fraud including benefit and tax crimes, crimes of negligence or financial exploitation. Many victims of crime live on the margins of society, homeless for example, where surveys struggle to reach.

The National Research Council of the National Academies in the United States convened a high-level workshop of criminologists and statisticians to consider such problems, the results of which were published in 2003. It concluded that:

> The clandestine nature of many crimes means that the victim may be unable to provide key details about the victimization and may not even be aware that a crime has been committed at all. Certain incidents that are supposed to be reported in an interview may seem irrelevant to respondents, since they do not think of these incidents as involving crimes. For example, victims of domestic violence or of sexual harassment may not think of these as discrete criminal incidents but as chronic family or interpersonal problems. It may be difficult to prompt the recall of such incidents with the short, concrete items typically used in surveys.

Such observations make clear that the practical challenges faced by the surveying envisaged here are substantial and that this may ultimately limit the range of crime to which this approach may be directly applied. At the same time, in estimating prevalence we are interested in only a count and are not limited to conventional face-to-face surveys of victims. For example, estimates of credit card fraud could incorporate cases identified by credit card companies.

Now let us turn to the third variable, P, precision.

3. Estimating Precision with a Gold Standard

In information retrieval, the simplest method of measuring the precision of an algorithm is against a set of results provided by a *gold standard* mechanism of human assessment.(Sometimes referred to in the legal literature as a 'blue riband', e.g. in Kaye, 1980, p. 1013) For example, if I have an algorithm for classifying books according to an indexing schema, I will ask an expert human to *code* by hand a sample of books. I will then ask the algorithm to classify the same books and construct a 4-gram by representing decisions that match the human *coder's* as true and the others as false. We can then calculate the precision and recall of the algorithm.(A good primer is Chapter 8 of Manning, Raghavan, & Schütze, 2008)

Such an approach relies on an assessment of the trustworthiness of the gold standard. Thus to pursue this approach it is necessary to believe that it is possible to find a way of determining guilt that will be more accurate than the system we are using to hand down verdicts. Or, conversely, to dismiss this approach it is necessary to believe that it is impossible to find such a way.

It is possible to make an argument for impossibility in the Epps vein. After all, one of the animating ideas of a trial is to bring to court all the relevant information, sift it thoroughly and so make the most accurate decision possible. Furthermore, the law strives to do this job as well as possible. How then can we expect an alternative system to do better?

First note that the ambit of the jury trial has been narrowing across many jurisdictions. In the Anglo-Saxon nations it has fallen into disfavour for civil cases and is often now by-passed by plea bargaining, magistrates courts or other mechanisms for processing guilt. On the continent of Europe, many jurisdictions that did rely on it have abandoned it. All sorts of reasons may lie behind this, including cost, but accuracy may also be part of it.

Further, in England and Wales, the government brought forward legislation, eventually defeated in 2007, to replace the jury in complex fraud cases with judges. Speaking for the government, the Attorney General said in the final debate in the House of Lords, '... serious fraud cases are distinct in that even an enormously long trial may not, for the reasons I have given, ensure that justice is done.' In other words, the government believed that judges are more capable of making an accurate determination in such cases than juries. Notwithstanding that these are a special kind of case, the point of principle stands: the accuracy of current arrangements may be improved upon. ('Fraud (Trials without a Jury) Bill, 2nd Lords Reading', 2007a)

Then there are several kinds of substantive reason that flow from the fact that a trial is not solely concerned with getting at the truth. Tribe emphasized the importance of ritual rather than accuracy as an organizing imperative. Stein cited Weinstein, who listed as competing

factors, '... economizing of resources, inspiring confidence, supporting independent social policies, permitting ease in prediction and application, adding to the efficiency of the entire legal system, and tranquilizing disputants.' In defeating the government's plans in 2007, Lord Kingsland claimed a role for trial by jury in buttressing democracy.('Fraud (Trials without a Jury) Bill, 2nd Lords Reading', 2007b; Stein, 2005, p. 36; Tribe, 1971; Weinstein, 1966)

One important non-epistemic concern is the inalienable rights of the defendant. Stewart has recently examined the distinction between such rights and contingent privileges conferred on defendants as part of the quest for accuracy and argues that the former includes, for example, the right to silence.(Stewart, 2016, p. 380)

Another, possibly even more telling non-epistemic concern, is the court's interest in punishing those who are convicted, the prompt for masses of lying, not only from defendants but also from witnesses.

Then there are the organizational arrangements for the trial. Some of these clearly have a down side in terms of accuracy; for example, holding the trial in public deters some witnesses from testifying. Some are more mysterious, starting with the adversarial system itself. Our adherence to this is not founded on evidence of its accuracy and Langbein has said we 'live under a criminal procedure for which we have no adequate theory.'(Langbein, 2003) The same applies to more detailed arrangements.

Finally, there are the policies that we adopt, including the standard of proof. Many of these are deliberately designed to reduce false convictions at the price of an overall loss in accuracy. Others have been developed to deal with the specific character of juries; for example, the exclusion of prejudicial evidence.

The purpose of a gold standard in the law is not to hand down individual verdicts and it does not need to be capable of fulfilling that function. Its purpose is only to assess at the gross, statistical level the accuracy of the verdicts handed down. To the extent that the policies of the criminal justice system are affected by non-epistemic concerns, so the scope for epistemic error is increased. Since there are so many of these concerns and their impact on our assessment may be large the scope seems likely to be large. By abandoning these other concerns and focusing solely on truth we can create space for an alternative gold standard assessment of guilt that we would indeed expect to be more accurate and which could be used to generate statistics through examination of a relatively small sample of cases.

If we try to sketch how a gold standard might operate, we see that we have many options for deviating from the current arrangements. We don't know yet how significant each of them is and there will always be a cost/benefit trade-off. But if we chose to take advantage of all of these possibilities, what we might end up with is this:

1. We establish an agency that is not part of the criminal justice system but is responsible for measuring its epistemic effectiveness, just as a statistics agency is not part of the education system but may be responsible for measuring its performance
2. It appoints assessors to a gold standard panel
3. A sample batch of cases in which a verdict has been reached that are reasonably similar (e.g. all rapes) are selected
4. The assessors re-investigate in private, including both re-questioning the defendant and questioning witnesses who did not testify in the case itself, in both cases after removing the threat of punishment and reviewing which of their inalienable-when-threatened-with-punishment rights should stand. This may or may not involve an adversarial framework.
5. Each assessor then returns their assessment of guilt independently.
6. Each element in the 4-gram is then composed of a comparison between a single assessor and the actual result of the case. From this we can calculate precision.

In step 5, assessing guilt, the assessors try their best to correctly decide whether the defendant truly committed the crime or not. To put this in jurisprudential terms, this means that the standard of proof for the assessment is the balance of probabilities. If one believes, as a result of magical thinking or subjective measurement, that the current standard of proof is effective in tilting the balance against false convictions, then it can be expected that many defendants will be classified as false negatives simply on the basis of the different standards of proof. Note however that: i) the number of false negatives is not involved in the computation of precision; ii) in most categories of crime and given the sovereign perspective, this variation will be swamped by false negatives arising from failure to identify, charge and prosecute.

This suggestion is just one approach to establishing a gold standard, the purpose of which is merely to illustrate that such an exercise is possible. Anyone with a practical turn of mind will immediately and justifiably wonder whether what is set out here would be ideal. For example, even though we now lack a defendant *per se*, why choose inquisitorial over adversarial? This is not a challenge I will attempt to meet. The critical question is whether a gold standard would really be more accurate than the system we already have. Even this less demanding question is something that cannot be definitively answered in advance. Without actually doing some measuring, plausibility is the only possible standard, which has been the point of this section. However, that does not mean that the question could not be resolved in future, and demonstrating this capacity is the purpose of the next section.

4. Reliability and Validity

How would we know whether the so-called gold standard was in fact more accurate than the existing jury process? This is a question that is not unique to the law. If a phenomenon can be known through measurement there are usually competing methods or instruments. It is a question Hand discusses under two headings, *reliability* and *validity*. Reliability says that the assessors should agree with each other, validity that they should not be systematically biased. The same concepts appear in many fields under a variety of labels. Let us look at the two in more detail.

Reliability may have several different dimensions but for our purposes can be narrowed down to the consistency of decisions between *coders* and can be thought of as invariance to changes in personnel. This is not to say that anyone could be a coder; rather, it is invariance within a well-defined group of coders, with the definition of that group being fundamental to the character and authority of the measurement.

A virtue of reliability as a concept is that it can be unambiguously measured. A common measure of such *inter-rater reliability* used in information retrieval is Fleiss's kappa, a variant of the better-known Cohen's kappa. Schaer *et al* provide a recent example of its usage. Scores can vary between 0 (no agreement) and 1 (all coders agree on everything) and, as with confidence intervals, conventions are emerging concerning the kind of language that should be used for different scores – e.g. the suggestion that scores between 0.6 and 0.8 represent 'substantial agreement'.(Schaer, Mayr, & Mutschke, 2010 Section 2.2)

An apparent argument against reliability is the deliberate inconsistency of jury nullification, which we can interpret generously as the reasonable interposing of the community's common sense between the law and the defendant. Reliability may be seen as a threat to this, but we need to be careful to distinguish two different kinds of consistency here: invariance to changes in personal and compliance with the dictates of the law. To see this, imagine having several juries considering the same trial and all handing down verdicts; these might all agree the defendant was guilty in law but should nonetheless be acquitted, making them consistent in the first way but inconsistent in the second. Thus we can see the two kinds of inconsistency are orthogonal concerns; we can therefore put jury nullification to one side and concentrate purely on inter-rater reliability.

This kind of reliability can provide evidence in support of the gold standard. We can measure the reliability of juries, e.g. by building a courtroom with space for more than one jury. We can also measure the reliability of the gold standard. And this provides a requirement for any gold standard: it should be more reliable than the juries it is assessing.

Validity can be thought of as a question of systematic bias.[3] Unlike reliability, it does not admit of a simple calculation. In discussing the validation of a measuring instrument (as kinds of which we can consider both the trial process and gold standard) Hand says:

> Perhaps the best one can hope for is a carefully argued process of attempting to establish validity, so that other potential users can agree that what has been done is solid. A process of 'due diligence', perhaps, or 'protocol, in the sense of Bird *et al*…

Thus to establish the overall accuracy of the gold standard, we need to combine consideration of both reliability and validity. Reliability we can straightforwardly measure. Validity requires the more rounded 'carefully argued process' that Hand describes. Such patient reasoning in which evidence of all kinds is commissioned, marshalled and considered is however familiar both to the law and government administration and should hold no terrors.

There is no magic possessed by the physical sciences that exempts them from this, and in that those who insist science is a social phenomenon are correct. There is always a question of trust. It is true that the physical sciences often find it relatively easy to establish measurements that are accurate to small fractions of a single per cent, a preciseness that the social sciences and the law cannot dream of matching. But this is a difference of degree rather than principle. Establishing one instrument (for example the gold standard) as more accurate than another (the trial system) is simply the everyday business of measurement in the modern world.

5. Confidence and Combinatorics

As an example of how this methodology could be used to determine policy, let us now examine a question that continues to vex jurisprudence, how to choose the standard of proof.(Quantitative attacks on this go back at least to Laplace, Poisson and Venn. See Venn, 1888, p. 328) Given that we have made the value judgement of b^2, the best standard in this framework is the one that maximises the F_b-measure. How can we find out what that is?

Rather than the continuously variable percentages popular in probabilistic approaches, let us adopt Risinger's 'well-ordered system of categories', known more generally as an *ordinal scale*. These could be 'beyond reasonable doubt', 'clear and convincing evidence' and 'on the balance of probabilities' though, as will become clear, there are other ordinal scales we may want to adopt. Then the *t*-axis of Figure 6 consists of this list with gaps of nothingness between its entries.(Risinger, 2017, p. 981)

[3] Psychometrics, the venue for much of the most sophisticated thinking about measurement, often terms it *veracity*.

We need measurements of precision and recall for the different points on the scale. Only once we have these can we determine which is the most effective option. There are at least three ways to obtain these.

Option 1 is to give the verdict-giver the option of specifying the category that reflects its degree of confidence. Scotland already provides an example of a jurisdiction in which this distinction is routinely made through the three allowable verdicts of guilty, not proven and not guilty. Further extending the available categories in a clearly ranked way would provide additional information and improve the resolution of our microscope. For separate reasons, both Laudan and Picinali have recently floated systems with multiple verdicts.(Laudan, 2010; Picinali, 2018b)

This path is not unproblematic however in that i) there is evidence that jurors in Scotland are confused about the three options they already have, let alone more; and ii) Lillquist's survey of several studies suggests jurors have a poor grasp of different standards of proof.(Hope, Greene, Memon, Gavisk, & Houston, 2008; Lillquist, 2002, p. 111)

We may then turn to Option 2. In this case we note that an ordering such as Risinger's suggests an underlying quality that is being ordered, an axis of some kind, and that in this case the obvious way to describe this is as the *confidence* that the jury requires to convict.

It may seem that the idea of a confidence threshold for a verdict and the standard of proof are synonymous, but this need not be the case. Instead of a jury of 12, consider instead a collection of 6 *micro-juries*, each composed of 3 jurors, each offering an independent *opinion* on the defendant in question. We can set what we consider to be an appropriate threshold for conviction – for example a guilty opinion from 5 of the 6 – and hand down the verdict to the defendant on that basis.[4] So long as we record the verdicts of all the micro-juries, what we have acquired along the way is a data set that can be used to model the performance of the system as confidence thresholds change. We can, as it were, re-run the algorithm with a threshold of, say, 6 guilties or 4 guilties and see what the impact is on outcomes.[5] [6]

By introducing this dimension of what I shall call *combinatorics*, we see that it is possible to separate the standard of proof, a written standard that human beings refer to, from

[4] There is a considerable discussion to be had about what an appropriate threshold would be.

[5] As an aside, micro-juries provide a simple means of managing the degree of reliability and representativeness of jury verdicts. More micro-juries implies more reliable and representative results in a demonstrable way that escapes larger juries and majority verdicts, wrapped up as they are in questions of group psychology. Distinct thresholds would also allow the introduction of a system of multiple verdicts as suggested recently by Laudan and Picinali. (Laudan, 2010; Picinali, 2018b)

[6] This may (or may not) leave us with the problem of reconciling the micro-juries ordinal scale with the written categories ordinal scale used by conventional juries, but this could be attacked through empirical calibration.

confidence in the guilt of the defendant. At the same time, we make it possible to obtain the observations we need.[7]

Option 3, a related approach, is to record the number of jurors voting in favour of guilt at the end of their deliberations. Then a vote of 7 in favour of guilt is a category above a vote of 6 in favour. True, the effect of the group dynamics in a jury is opaque. But if we do not think that 12 votes in favour provides more confidence than 11, and that more than 10, then there is no logical basis for any kind of threshold at all, including unanimity.

These techniques can be used to address all epistemic policy questions. There are some questions where it is clear that we can reduce errors of all kinds – DNA profiling for example. In such cases the path of policy is clear. The others are *epistemic policy dilemmas* in which we would like to reduce false convictions but fear increasing false acquittals. Through these we can shift the burden of proof one way or the other as depicted by Dahlman. Thus we plot the curve of the F_{β}-measure as in Figure 6 using one of the ordinal scales set out here and look to see whether we are currently sitting to the left or right of the maximum. From this we know whether we need to shift the burden of proof one way or the other. It does not tell us what the ideal mix of policies is, a specially complicated question as they are not independent, but it does provide an iterative procedure through which, as in information retrieval, we can continually improve performance.

Combinatorics assumes a heightened importance in calculating precision if we consider representativeness more important than accuracy as the ultimate source of legitimacy for the system, a distinction that we can clarify with reference to Kaye's analysis of jury quality.

In analysing arguments over the minimum size of juries in the United States in *Ballew vs. Georgia*, Kaye parsed the concerns into three: "Inasmuch as the 'quality' of jury performance is an amorphous term, the constitutional inquiry should be focused on more specific components: representativeness, accuracy, and reliability." (*Ballew v. Georgia*, 1978; Kaye, 1980, p. 1013) 'Reliability' Kaye uses exactly as we do here, but 'accuracy' for him is equivalent to 'validity' for us. The odd one out therefore is representativeness, which in itself is not part of our conception of accuracy at all. Is it a completely independent concept, or related in some way to the other two?

Representativeness may have two distinct meanings. One is that on certain parameters, the proportion of jurors matches the proportion in the population at large – men/women,

[7] The above discussion suggests that we should parse the verdict-giving function into six elements: the *segmentation* into distinct verdict-giving units; the *personnel* such as judges or lay people; the *available verdicts*; the standard of proof; the *opinion-gathering procedure* by which each independent unit arrives at a decision, for example by unanimity or simple majority; and the combinatorics, the way in which the decisions of the independent units result in a verdict. A seventh concern is the kind of *information* about the innards of the process that is recorded and revealed in what ways to whom.

whites/blacks and so on. This is the way it is usually discussed but is not at all an aspect of *performance*, Kaye's concern and mine; that is an input rather than an output. For the jury's performance to be representative, the decision itself must match that of the population (which can only mean here 'what the population would have concluded if it had sat through the trial').

The binary nature of a verdict obscures this somewhat. How can the decision of a jury be representative when there is less than unanimous agreement in the population? If we imagine instead verdicts returned as a number representing the jury's confidence in the defendant's guilt things get clearer. Then a jury might say its confidence was 50 per cent and, if the population at large was evenly split on guilt, we could say that the jury was representative.

To rely on representativeness is to accept whatever systematic biases there may be in the population. If the population is racist, then a representative jury system will be racist. Thus it is fundamentally disconnected from validity. With regards to reliability, we can consider two questions. First, can an unreliable trial system be representative? Yes, verdicts in individual cases could vary wildly but on average the results could be reflective. Second, can an unrepresentative trial system be reliable? Yes, we could restrict juries on offences involving cats to cat haters who were nonetheless consistent in their hatred of cats and hence in their verdicts. Thus there is no automatic connection with reliability either. Kaye is right, representativeness is a distinct, third concern.

To the extent that we think that representativeness is important, we can justify calculating precision in another way, *pairwise*. This is an approach already used in, for example, clustering analysis. In the legal context it could work like this. Start with a trial with 6 micro-juries and treat each of the 6 in turn as the gold standard. For each one, compare its verdict with the verdict of each of the other 5 and classify the result as a TP, FP, TN or FN. We could write each result in the form (3, 5) = TN, meaning that when we take the 3rd jury as the as gold standard and assess the 5th jury by it, the result is a true negative – i.e. both returned not guilty verdicts. We now have a universe with 6 * 5 = 30 elements in it and can draw the 4-gram.

The total number of false positives must be the same as the total number of false negatives because for every pair where the first verdict is guilty and the second innocent, there must be another pair where the first verdict is innocent and the second guilty. For example, if (2, 5) = FP then (5, 2) = FN. By contrast, if (2, 6) = TP then it must also be the case that (6, 2) = TP; no TN comes into it, and so there is no reason why the tally of true positives should be the same as the tally of true negatives.

If we aggregate the pairwise data from many trials and calculate the precision yielded by this universe, what we are getting is an indication of the inconsistency of verdicts assessed

against the yardstick of agreed guilty verdicts. If representativeness rather accuracy is your bedrock, it is a viable definition of precision.

6. Worked Example of Policymaking via the M-N-P Computation

We can now illustrate the M-N-P method of computing the F_{β}-measure and so making a policy decision with an example. The purpose of this is merely to illustrate the steps in the computation. It is not an attempt to model the reality with reasonable assumptions, which would be undermined by the fact that none of us knows how precision and recall vary with confidence.

The text in this section is an annotation to the data shown in the figures. The basic scenario is that we are evaluating precision and recall in cases of rape where decisions are made by a set of 6 micro-juries with a threshold of 4 guilty opinions required for conviction. We have selected 10 cases for assessment by the gold standard.

Figure 7 shows the opinions of the micro-juries, labelled A to F in each case. G = Guilty; NG = Not Guilty. Cases 6, 7, 8, 9 and 10 resulted in a guilty verdict being handed down to the defendant; the other five resulted in a not guilty verdict.

Figure 7: Decisions by micro-juries

	A	B	C	D	E	F		
Case							Total Guilties	Verdict
1	NG	NG	NG	NG	NG	NG	0	Not Guilty
2	NG	NG	G	G	NG	NG	2	Not Guilty
3	G	NG	NG	NG	NG	NG	1	Not Guilty
4	G	G	G	NG	NG	NG	3	Not Guilty
5	G	G	G	NG	NG	NG	3	Not Guilty
6	G	G	G	NG	G	NG	4	Guilty
7	G	G	G	NG	G	NG	4	Guilty
8	G	G	G	NG	G	G	5	Guilty
9	G	G	G	NG	G	G	5	Guilty
10	G	G	G	G	G	G	6	Guilty

Figure 8 then shows the rest of the computations, which are all organized according to the 7-point ordinal scale implied by having 6 micro-juries.

Figure 8: Computation of the $F_{1/\backslash 10}$-measure

Ordinal scale of confidence		**0**	**1**	**2**	**3**	**4**	**5**	**6**
1. Verdicts by threshold								
Case	Total Guilties							
1	0	G	NG	NG	NG	NG	NG	NG
2	2	G	G	G	NG	NG	NG	NG
3	1	G	G	NG	NG	NG	NG	NG
4	3	G	G	G	G	NG	NG	NG
5	3	G	G	G	G	NG	NG	NG
6	4	G	G	G	G	G	NG	NG
7	4	G	G	G	G	G	NG	NG
8	5	G	G	G	G	G	G	NG
9	5	G	G	G	G	G	G	NG
10	6	G	G	G	G	G	G	G
2. Assessment by Gold Standard								
Case	Assessor's Verdict	Epistemic Outcomes						
1	NG	FP	TN	TN	TN	TN	TN	TN
2	NG	FP	FP	FP	TN	TN	TN	TN
3	NG	FP	FP	TN	TN	TN	TN	TN
4	G	TP	TP	TP	TP	FN	FN	FN
5	NG	FP	FP	FP	FP	TN	TN	TN
6	G	TP	TP	TP	TP	TP	FN	FN
7	NG	FP	FP	FP	FP	FP	TN	TN
8	NG	FP	FP	FP	FP	FP	FP	TN
9	G	TP	TP	TP	TP	TP	TP	FN
10	G	TP	TP	TP	TP	TP	TP	TP
3. Computation of P (Precision)								
TPs		4	4	4	4	3	2	1
FPs		6	5	4	3	2	1	0
Precision		0.40	0.44	0.50	0.57	0.60	0.67	1.00
4. Computation of M (Convictions)								
Raw (TPs + FPs)		10	9	8	7	5	3	1
Scale factor	*160.4*							
Scaled		1604	1444	1283	1123	802	481	160
5. Computation of Recall								
M		1604	1444	1283	1123	802	481	160
N		69000	69000	69000	69000	69000	69000	69000
P		0.40	0.44	0.50	0.57	0.60	0.67	1.00
Recall		0.0093	0.0093	0.0093	0.0093	0.0070	0.0046	0.0023
6. Computation of $F_{ß}$-measure								
Precision		0.40	0.44	0.50	0.57	0.60	0.67	1.00
Recall		0.0093	0.0093	0.0093	0.0093	0.0070	0.0046	0.0023
b^2	*0.10*							
$F_{ß}$-measure		0.083	0.085	0.086	0.088	0.069	0.048	0.025

In Step 1, the different possible thresholds for conviction are considered and the verdicts that would have been handed down in each case are tabulated.

Step 2 makes the comparison with the gold standard. For simplicity, I have only introduced the assessments of one assessor, which in aggregate are four guilty and six not guilty. On the basis of these, the verdicts from the micro-juries are classified as true negatives and so on.

In Step 3, the total number of true positives and false positives are added up and precision calculated.

In Step 4, first the number of convictions *in the sample of 10* are calculated. Then we introduce some real-world data.

The only contemporaneous figures for prevalence and conviction for a well-defined category of offences that I can find is a report into sexual offences in England and Wales published in 2013, specifically the figures for rapes committed in 2009 (although even these are imperfect). These resulted in 802 convictions from an incidence estimated at 69,000.[8]

The 802 convictions correspond in our scenario to convictions at the *de jure* threshold of 4 micro-juries. Thus the 5 convictions from our sample of 10 stand for those 802. As the number of convictions in our sample of 10 increases or decreases, so the number of convictions in the real world should be expected to increase or decrease. Thus there is a scale factor of 160.4 and we multiply the raw figures from the survey by this to generate the appropriate, scaled value.

In Step 5, recall is computed with the M-N-P method.

In Step 6, we adopt as our value Blackstone's ratio, $b^2 = 1/10$, and then compute the $F_{1/\sqrt{10}}$-measure.[9]

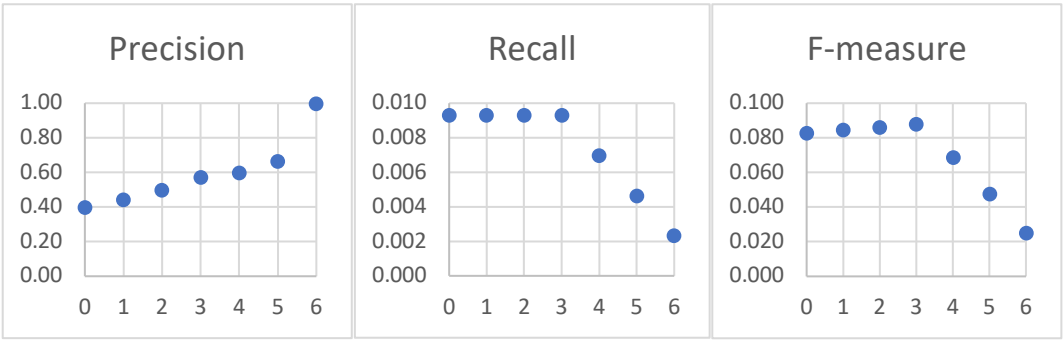
The results of these computations are then shown in the charts in Figure 9. As precision rises, recall falls. The $F_{1/\sqrt{10}}$-measure peaks at the threshold on the scale of 3 micro-juries. Therefore, given our value judgement of the relative importance of false convictions and false acquittals, we should reduce the threshold for conviction from 4 to 3 micro-juries. Alternatively, we could adjust other policies that shift the burden of proof in favour of the prosecution.

---

[8] *An Overview of Sexual Offending in England and Wales*, Statistics bulletin, Ministry of Justice, Home Office & the Office for National Statistics, 10 January 2013, 10 and Table 2.2, 13. Even here issues with both data and method remain to be resolved. The figure for convictions is for offences committed in 2009 but that for prevalence is an average of 2009-12. The figure for convictions is only that for rape and not other, lesser offences. No allowance has been made for perpetrators who died before they could be brought to trial. It is not clear if any of the convictions was for more than one rape; or the extent to which the general prevalence involves offences with more than one perpetrator.

[9] I adopt Blackstone's ratio as an unavoidable *de facto* reference point in order to try and avoid creating my own Overton window.

Figure 9: Charts of Precision, Recall and the $F_{1/\backslash 10}$-measure



7. Conclusion

The main purpose of this article has been to establish that we can obtain the formal observations of the epistemic outcomes of the criminal justice system that will allow coherent computation with the F_β-measure resulting in objective rather than subjective measurement of effectiveness. The answer I have provided at the very highest level is: we count convictions; we estimate crimes by survey; we estimate precision by a gold standard; and using the M-N-P method we then calculate the F_β-measure. I have sketched ways in which we might go about getting the estimates and I have shown how this approach can be applied to the making of policy through the example of the standard of proof.

The measurements obtained in this way can be trustworthy and hence would provide a new, objective basis for deciding between policies and so determining the operations of the entire criminal justice system. Blackstone's ratio, which has fallen out of favour in many parts of the Anglo-Saxon world in recent decades, would be restored to prominence, but in a properly consequential way.

All of this is based on the primacy of outcomes. Envisaged is a re-organisation of the law in which the procedural or ritualistic is downgraded and instead orchestration is primarily directed at maximizing the epistemic good as defined by b^2. The choice of b^2 involves consideration of the many crimes that never reach the police, let alone the courts. Thus it is not a question of jurisprudence or law or for any institution of the criminal justice system. It is a kind of interface between the sovereign, or the popular will, and the system that is in my view so consequential as to be a matter for constitutions. To make the case for this new emphasis on outcomes I will advance six arguments.

First, justice is generally considered by most people for most things a question of outcomes, and the law is not immune to this. Los Angeles in 1992 did not descend into riots when the jury to try the police officers who beat Rodney King was formed without a black on it. The riots started when the jury failed to convict the defendants.(Mydans, 1992)

Second, epistemic measurement is already deeply embedded in the heart of our legal system but has succumbed to the kind of invisibility that Hand says is common.(Hand, 2004, p. 1) To make it visible, let us look backwards.

The jury is a very old part of our legal furniture and some kind of explanation is required for why it has survived when other aspects, trial by battle for example, have not. I don't want to attempt a complete answer to that question, but instead to examine it narrowly through the lens of two characteristically modern values, the quantitative and the democratic. I will do this with the help of the woodcut in Figure 3 setting out how to establish the length of a foot. This comes from the early modern period in which the current legitimation of the jury was forged, when the idea of man as the measure of all things was not an abstraction but a physical feature of daily life throughout Europe.

To today's eyes this procedure may look unreliable, but Koebel's book marked a substantial advance in understanding and has stood the test of time. At a time when units of measurement routinely varied from town to town, this process of random sampling and averaging would have yielded a local standard that was reasonably consistent and trustworthy and hence a viable basis for commercial transactions.

Taking 12 men and asking for a verdict could therefore be seen as a means of generating a standard of measurement *in the same way*. One similarity is that the composition of the group is supposed to be of no significance. A second, related similarity is that the groups should be interchangeable. These are both aspects of reliability. In short, one fundamental reason for our faith in the jury is that we recognise that its structure – and hence its measurements and outcomes – is intrinsically reliable. This is a quantitative virtue.

Then consider two further similarities. First, the group is supposed to be representative of society at large, the part standing for the whole; specifically, measurements based on the part should be the same as those yielded by the whole. Second, the public and random selection of the men, a kind of impartiality, is part of establishing the authority of the measure; just as a wooden stick might be forged, so an appointed verdict-giver might be corrupted.

These two features are also found in sortition, a form of government that dates back to Antiquity and was a feature of Europe before and during the period in which the jury has its origins.[10] They are aspects of measurement that, when used to make decisions of the state, are democratic virtues.

[10] Sortition in government means not having elections and instead appointing legislators or the executive through a lottery. The government formed in such a way is representative and democratic but not a "representative democracy" in the electoral form in which we know it. Popular in the Italian city states, it continued in Venice up into the late eighteenth century. (DOWLEN, 2009, PP. 1, 67)

So while modern ideas of the quantitative and democratic would in the coming centuries overthrow many traditions in the law as elsewhere, the jury was not at risk because it already incorporated sound quantitative and democratic principles that are aspects of measurement: reliability, impartiality and representativeness. These remain virtues today.

Incorporation of these principles in the mechanics of the jury is however not the same as measuring the epistemic outcomes of the system as a whole. This has not been possible because we have lacked until recently both a suitable computational framework and the institutional capacity to estimate precision and recall. That inability has prevented the law from focusing on outcomes, but once it is overcome the emphasis on procedure needs to be reconsidered – the third reason.

Fourth, the central objection to this new approach I suspect is likely to be a fear that it would undermine the law. It is striking that no jurisdiction has yet seen fit to measure the reliability of jury verdicts. More compelling than the desire for measurement's insight, it seems, is fear of lost legitimacy, vividly expressed by Tribe in 1971 when he suggested that the union of mathematics and the trial process '… would be more dangerous than fruitful.' (Tribe, 1971, p. 1393) While understandable, this fear would, I think, be misplaced today.

One simple riposte to this fear can be found in a recent speech by Lord Neuberger, a Justice of the UK's Supreme Court: 'The fact that we cannot get the answer right every time is no excuse for not doing our best to get the right answer.'(Neuberger, 2016)

It must be acknowledged, however, that through measurement we would certainly clarify the persistent, structural existence of error. We could expect, for example, annual figures of precision and recall for rape. This would break with the vagueness of the discussion engendered by the current, subjective measurements and be at odds with the majority opinion in *In re Winship,* the decision that has given beyond reasonable doubt the force it has in the United States, which concluded that, '[i]t is critical that the moral force of the criminal law not be diluted by a standard of proof that leaves people in doubt whether innocent men are being condemned.'(*In re Winship*, 1970)

Thus one might conclude that measurement would indeed dangerously undermine the legitimacy of the law. But is it really that simple? Since *In re Winship* and Tribe we have had, across the liberal democracies of the West, half a century of increasing public exposure of the law's failings.(Darbyshire, 2015) Protestations of perfection are surely beyond credibility today. In pursuing such claims now, would the law not simply convince the world that it is not to be trusted? Is it not more trustworthy in 2018 to start by accepting the evident nature of our predicament, to keep it real?

If one reads Neuberger's speech and his frank admissions of error, it might seem that this argument has already been won, in the UK at least. But he is addressing judicial error and we

still have no quantification. If we have matured to the point where we can acknowledge systematic epistemic error in the criminal justice system, we have done so without realising the potential this provides for reducing error and increasing legitimacy, the two potential benefits of measurement's insights to which I will now turn.

Fifth, part of this approach, as exemplified by the standard of proof, is about shoving policies one way or the other, not blindly but deliberately via a choice of b^2. But another part of it is engineering the system to be more accurate regardless of b^2, a difficult task in the absence of measurement. As an example, consider that even a perfectly valid system may return many errors of both types if it is unreliable. Increasing reliability throughout the system could yield improvements that far outstrip, for example, the benefits of DNA profiling. Micro-juries offer this potential in a way that increasing the number of jurors on a conventional jury does not, and measurement could confirm this.

Sixth and finally, objective measurement of outcomes promises a new and distinctively modern source of legitimacy.

Although complicit in it, for example through the allocation of its budget, the sovereign is not part of the criminal justice system. It looks at the system from outside, as a customer might look at a contractor. The contractor has a job; is it doing it well or badly? Thus the choice of whether to focus on outcomes is not one that is properly offered to the institutions of the law. It is up to the sovereign to tell the system what it wants and to decide by which criteria it will judge the system.

To an extent, the state already does judge the system on outcomes. It could not tolerate a system that was, for example, epistemically random or destroyed recall by insisting on perfect precision or destroyed precision by insisting on perfect recall. Such subjective measurements ensure the system is adequate from the state's point of view in terms of its outcomes; and, if it is not, it will be altered until it is, for example through the introduction of plea bargaining or other measures for processing guilt. However, without objective measurement of outcomes of the kind proposed here, there is a democratic deficit, which I will try to delineate by considering Blackstone's dictum that '…it is better that ten guilty persons escape than that one innocent suffer'. (Blackstone, n.d. *359 (first published 1765, Lippincott Company 1893).)

Despite its pervasive influence, we have no history of the dictum – where it came from, how it achieved such a central position, the role it played over the centuries, a narrative of its decline from incontestable. Without this it is difficult to see the size and shape of the hole it has left behind. My own view of its ascent depends on first restoring the dictum's immediate context in the *Commentaries* as the first step in a larger, three-step argument of justification for certain policies, and then on Shapiro.

Shapiro has described how the law overcame the crisis of proof when it lost access to divine guidance through the medium of the jury's Christian conscience. Through the development of evidential rules, she argues, the law persuaded the public that the jury retained the 'divine spark' and could be relied upon to deliver accurate verdicts.(Shapiro, 1993, p. 241) From this, I have reasoned as follows to explain the dictum's original popularity:

> The explanation seems to me to likely lie in a combination of four factors: the new need for a justification for policies identified by Shapiro; the usefulness therefore of the very general applicability of steps one and two in Blackstone's argument; the dictum's quantitative aspect, which chimed with the beginnings of the general adoption of the quantitative in society at large, a movement which at about the same time gave us Beccaria's *la massima felicità divisa nel maggior numero*; and the ability of the dictum to be read as shoving policy-makers towards policies that reduce both the number of wrongful convictions and the ability of the state to seize and convict, which chimed with the new and popular demands of the emerging democracies and was in contrast with other contemporary frameworks, such as that noted by Franklin which invested judges under Robespierre with an absolute discretion.
>
> The dictum therefore can be seen as a new kind of buttress of the law that was required in a new kind of society. By virtue of the 10 and the clear analytical framework it suggests of four objectively distinct outcomes, it partakes of the quantitative; by virtue of this apparent transparency and the shove, it partakes of the democratic; and these two aspects together give it one foot in modernity. But the lack of a clear and cogent argument, of rationality, and the lack of a mechanism of policy-making susceptible to democratic oversight leave it with one foot in the pre-modern.

From this viewpoint, the nature of the hole left by the dictum's decline becomes clear. The attempt to step into the modern has failed and in addressing the many profoundly consequential epistemic policy dilemmas the law has been thrown back into an essentially pre-modern state, devoid of the quantitative and democratic virtues the dictum promised but failed to deliver. Thus the incoherence in this area of law identified by Hamer. Thus the turn in recent decades to purely moralised accounts, such as that begun by Duff. Thus also the failure, recently noted by Picinali, of any of the currently competing attempts – including consequentialism, deontological theory, and retributivism – to adequately justify a standard of proof on anything but its own terms; since all the contending theories lack the quantitative wires along which democratic consequence and legitimacy could flow, the debate is

incapable of being resolved in a democracy.(Duff, 1986 Chapter 4. Hamer, 2004; Picinali, 2018a)

While general, the failure of reasoning described here is uniquely exposed in the United States through explicit engagement with the quantitative in i) the Supreme Court's work on *In re Winship* and *Ballew vs. Georgia*; and ii) the continuing practice whereby high courts in 38 states have adopted numerical values for the ratio of false convictions to false acquittals.[11]

Given the pages of calculation in this article and its preoccupation with measurement, it would be easy to see this as an attempt at a "scientific" approach to the criminal justice system. I make no such claim and find the concept of science unhelpful. Rather, the strength of the approach set out here is that it makes good on the promise we found in Blackstone's dictum, providing the quantitative and democratic virtues we always hoped for. It makes outcomes the central issue. Through a choice of b^2 our society can define the good in exactly the way found in Blackstone. Then measurement, comprising the methodology of the F_β-measure and the acquisition of observations as described here, makes that definition consequential for the first time. The law (and all its agencies) is given clear guidance on how to organize itself, and its performance can be scrutinized, just as we scrutinize universities or hospitals.

To go this way is to trade one kind of legitimacy for another, the pre-modern for the modern. It is to embrace rather than resist the process Habermas described in which '…scientific progress undermines traditional legitimating myths and forces the state to increasingly rely on science as an apparently-neutral basis for political decisions…' – albeit with the important deviation that the need to make a consequential choice of b^2 for the first time means that the effect of this is the exact opposite of the de-politicisation Habermas assumed.(Habermas, 1971, pp. 81–85) Measurement makes the body of the law subject to the sovereign's panopticon, with the kind of consequences Foucault describes.(Foucault, 1979) Control, authority and legitimacy run up and down quantitative wires. Through the technology of trust described by Porter, the current democratic deficit is eliminated. And, as with IQ tests, company accounts and public engineering, this is something done not *by* the law but *to* it.(Porter, 1996)

The costs of measurement may or may not turn out to be high, but we can make a comparison with other kinds of measurement systems such as the census, the compilation of economic statistics and clinical trials for pharmaceuticals. The billions spent on these are not

[11] The exposure may be explained by a cultural difference between the US and the UK in which, Porter suggests, America goes further in seeking to eliminate the scope for subjective reasoning; its accounting rules, for example, are far more prescriptive. (Pi, Parisi, & Luppi, 2018; Porter, 1996)

regarded as a waste. On the contrary, the measurements they provide are a necessary precursor to good decisionmaking, the clarity they provide essential to improving the trustworthiness and effectiveness of large and expensive systems that are of great importance to us all.

References

Ballew v. Georgia. , 435 U.S. 223 (1978).

Blackstone, W. (n.d.). *Commentaries on the Laws of England* (Vol. 4).

Conger, G. P. (1927). *Whitehead lecture notes: Seminary in Logic: Logical and Metaphysical Problems* (Yale University). Retrieved from Manuscripts and Archives, Yale University Library.

Crime in England and Wales - Office for National Statistics: year ending March 2017. (2017). Retrieved 11 October 2018, from https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/crimeinenglandandwales/yearendingmar2017

Cullerne Bown, W. (2018). The criminal justice system as a problem in binary classification. *The International Journal of Evidence & Proof*, *22*(4), 363–391. https://doi.org/10.1177/1365712718795548

Cullerne Bown, W. (2019). Killing Kaplanism: Flawed methodologies, the standard of proof and modernity. *The International Journal of Evidence & Proof*. https://doi.org/10.1177/1365712718798387

Dahlman, C. (2018). Determining the base rate for guilt. *Law Probability and Risk*, *17*. https://doi.org/10.1093/lpr/mgx009

Darbyshire, P. (2015). *'British justice is the finest in the world': An examination of Anglo-American boasting*. Presented at the In: Society of Legal Scholars (SLS) Annual Conference, York. Retrieved from https://eprints.kingston.ac.uk/33454/3/Darbyshire-P-33454.pdf

Doll, R., & Hill, A. B. (1956). Lung Cancer and Other Causes of Death in Relation to Smoking. *British Medical Journal*, *2*(5001), 1071–1081.

Dowlen, O. (2009). *The Political Potential of Sortition: A study of the random selection of citizens for public office*. Imprint Academic.

Duff, R. A. (1986). *Trials and Punishments*. Cambridge University Press.

Epps, D. (2014). The Consequences of Error in Criminal Justice. *Harvard Law Review*, *128*, 1065.

Foucault, M. (1979). *Discipline and punish: The birth of the prison. (Trans A. Sheridan)*. In *Discipline and Punish: The Birth of the Prison. (Trans A. Sheridan)*. Oxford, England: Vintage.

Franklin, J. (2002). *The Science of Conjecture: Evidence and Probability before Pascal* (Annotated edition). Baltimore: The Johns Hopkins University Press.

Fraud (Trials without a Jury) Bill, 2nd Lords Reading. (2007a, March 20). *Hansard*, p. Column 1149.

Fraud (Trials without a Jury) Bill, 2nd Lords Reading. (2007b, March 20). *Hansard*, p. Column 1152.

Habermas, J. (1971). *Toward a Rational Society: Student Protest, Science, and Politics* (J. J. Shapiro, Trans.). Boston: Beacon Press.

Hamer, D. (2004). Probabilistic Standards of Proof, Their Complements and the Errors that are Expected to Flow from Them. *University of New England Law Journal*, *13*, 221–242.

Hand, D. J. (2004). *Measurement Theory and Practice: The World Through Quantification*. London: Wiley-Blackwell.

Hand, D. J. (2016). *Measurement: A Very Short Introduction*. In *Very Short Introductions*. Oxford, New York: Oxford University Press.

Hart, H. L. A. (1973). Bentham and the Demystification of the Law. *The Modern Law Review*, *36*(1), 2–17. https://doi.org/10.1111/j.1468-2230.1973.tb01350.x

Hope, L., Greene, E., Memon, A., Gavisk, M., & Houston, K. (2008). A third verdict option: exploring the impact of the not proven verdict on mock juror decision making. *Law and Human Behavior*, *32*(3), 241–252. https://doi.org/10.1007/s10979-007-9106-8

In re Winship. , 397 U.S. 358 (1970).

Kaplan, J. (1968). Decision Theory and the Factfinding Process. *Stanford Law Review*, *20*(6), 1065–1092. https://doi.org/10.2307/1227491

Kaye, D. (1980). And Then There Were Twelve: Statistical Reasoning, the Supreme Court, and the Size of the Jury. *California Law Review*, *68*(5), 1004. https://doi.org/10.2307/3480278

Langbein, J. H. (2003). *The Origins of Adversary Criminal Trial*. Oxford University Press. (LAW D 25 SIM).

Laudan, L. (2006). *Truth, Error, and Criminal Law: An Essay in Legal Epistemology*. Cambridge University Press.

Laudan, L. (2010). Need Verdicts Come in Pairs? *The International Journal of Evidence & Proof*, *14*(1), 1–24. https://doi.org/10.1350/ijep.2010.14.1.338

Lillquist, E. (2002). Recasting Reasonable Doubt: Decision Theory and the Virtues of Variability. *University of California–Davis Law Review*, *36*, 85–197.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. New York: Cambridge University Press.

Medelyan, O. (2009). *Human-competitive automatic topic indexing* (University of Waikato). Retrieved from http://hdl.handle.net/10289/3513

Mydans, S. (1992, April 30). THE POLICE VERDICT; Verdict Sets Off a Wave of Shock and Anger. *The New York Times*. Retrieved from https://www.nytimes.com/1992/04/30/us/the-police-verdict-verdict-sets-off-a-wave-of-shock-and-anger.html

Neuberger, D. (2016). The Role of the Judge: Umpire in a Contest, Seeker of the Truth or Something in Between? Opening Remarks. *Singapore Panel on Judicial Ethics and Dilemmas on the Bench*. Retrieved from https://www.supremecourt.uk/docs/speech-160819-04.pdf

Packer, H. (1968). *The Limits of the Criminal Sanction*. Stanford University Press.

Pi, D., Parisi, F., & Luppi, B. (2018). *Quantifying Reasonable Doubt* (SSRN Scholarly Paper No. ID 3226479). Retrieved from Social Science Research Network website: https://papers.ssrn.com/abstract=3226479

Picinali, F. (2018a). Can the Reasonable Doubt Standard be Justified? A Reconstructed Dialogue. *Canadian Journal of Law & Jurisprudence*, *31*(2), 365–402. https://doi.org/10.1017/cjlj.2018.17

Picinali, F. (2018b). Do Theories of Punishment Necessarily Deliver a Binary System of Verdicts? An Exploratory Essay. *Criminal Law and Philosophy*, *12*(4), 555–574. https://doi.org/10.1007/s11572-017-9440-y

Porter, T. M. (1996). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press.

Risinger, D. M. (2006). Innocents Convicted: An Empirically Justified Factual Wrongful Conviction Rate. *Journal of Criminal Law and Criminology*, *97*, 761.

Risinger, D. M. (2008). Tragic Consequences of Deadly Dilemmas: A Response to Allen and Laudan. *Seton Hall Law Review*, *40*, 991.

Risinger, D. M. (2017). Leveraging Surprise: What Standards of Proof Imply That We Want from Jurors, and What We Should Say to Them to Get It. *Seton Hall Law Review*, *48*, 965.

Schaer, P., Mayr, P., & Mutschke, P. (2010). Implications of Inter-Rater Agreement on a Student Information Retrieval Evaluation. *ArXiv:1010.1824 [Cs]*. Retrieved from http://arxiv.org/abs/1010.1824

Shapiro, B. J. (1993). *Beyond Reasonable Doubt and Probable Cause: Historical Perspectives on the Anglo-American Law of Evidence*. University of California Press.

Stein, A. (2005). *Foundations of Evidence Law*. Oxford, New York: Oxford University Press.

Stewart, H. (2016). Concern and Respect in Procedural Law. In *The Legacy of Ronald Dworkin*. Oxford University Press.

Tribe, L. H. (1971). Trial by Mathematics: Precision and Ritual in the Legal Process. *Harvard Law Review*, *84*(6), 1329–1393. http://dx.doi.org/10.2307/1339610

Van Rijsbergen, C. J. (1979). *Information retrieval*. Retrieved from http://www.dcs.gla.ac.uk/Keith/Chapter.7/Ch.7.html

Venn, J. (1888). *The Logic of Chance an Essay on the Foundations and Province of the Theory of Probability, with Especial Reference to its Logical Bearings and its Application to Moral and Social Science, and to Statistics*. Macmillan.

Weinstein, J. B. (1966). Some Difficulties in Devising Rules for Determining Truth in Judicial Trials. *Columbia Law Review*, *66*, 223.