# PERFORMANCE CLAIMS IN FORENSIC SCIENCE EXPERT OPINION EVIDENCE

CHLOE A SMITH* AND MATTHEW B THOMPSON†

*In order for fact-finders to rationally evaluate the probative value of forensic pattern-matching evidence, information about the accuracy and reliability of examiners' opinions is necessary. Empirical tests of ability and performance, however, must first begin with a claim about what that ability and performance might be. In this article, we attempt to identify performance claims made by forensic pattern-matching disciplines by surveying professional literature published by representative discipline organisations in fingerprints, footwear and tyres, firearms and toolmarks, and handwriting and documents. Amongst these disciplines we did not find performance claims that are readily amenable to empirical testing. To spur progress, we suggest a basic framework to guide forensic disciplines toward formulating empirical claims that lend themselves to scientific testing: stipulate (1) the task you can perform, (2) the necessary conditions of performance, and (3) the standard of performance you can achieve. Once empirical claims are made, empirical tests can be designed and conducted that will help to fortify the scientific evidence base for forensic pattern-matching techniques.*

## I INTRODUCTION

Expert opinions provided by forensic examiners are used to help establish the facts of a case. Forensic science opinion evidence, with the exception of DNA, has gone largely unchallenged in court since its inception.[1] Several recent authoritative reports, however, have questioned the epistemic claims made by forensic examiners, and have highlighted the frequent absence of solid scientific research demonstrating the validity, reliability and accuracy of forensic analyses.[2] These reports, along with commentary in the broader scientific

* Murdoch University.

† Murdoch University. Corresponding author. Discipline of Psychology, Murdoch University, Perth, Australia. mbthompson@gmail.com.

1 Gary Edmond et al, 'Admissibility Compared: The Reception of Incriminating Expert Evidence (i.e., Forensic Science) in Four Adversarial Jurisdictions' (2014) 3 *University of Denver Criminal Law Review* 31; Gary Edmond, 'Latent Science: A History of Challenges to Fingerprint Evidence in Australia' (2019) 42(3) *University of Queensland Law Journal* (advance).

2 National Research Council of the National Academy of Sciences, *Strengthening Forensic Science in the United States: A Path Forward* (Report, 2009) ('NAS Report'); President's Council of Advisors on

community,[3] include recommendations for the urgent development of quantifiable measures of human performance in forensic pattern-matching.

Legal scholars have argued that in order for fact-finders to rationally evaluate forensic evidence they need information about the validity and reliability of forensic science techniques, including the limitations, proficiency and indicative error rates of forensic examiners' conclusions.[4] In this article, we contend that in order to fulfil the need for empirical testing of human matching-performance, and to provide this information to fact-finders, we first need to know what performance claims are being made by forensic examiners. Without reasonable and precise claims about ability and levels of performance, empirical studies cannot be designed and conducted. We will attempt to illuminate the claims made by forensic examiners by surveying the accessible professional literature across four forensic disciplines. We will then interpret these claims in terms of their amenability to empirical testing, and suggest a path forward for the forensic pattern-matching disciplines.

## II Background

In the past decade, several authoritative reports have emerged that describe a situation in which very little is known about the validity and reliability of forensic science evidence. In 2009, the United States National Academy of Science ('NAS') issued a report highlighting the absence of scientific evidence underpinning the forensic disciplines and the evidence provided by expert witnesses in court.[5] A Scottish Inquiry and a United States National Institute of Justice Report both highlight the inevitability of human error in forensic examinations.[6] A 2016 report from the United States President's Council of Advisors on Science and Technology[7] ('PCAST') reiterated the concern that forensic science lacks a strong

Science and Technology, *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods* (Report, 2016) ('PCAST Report'); National Institute of Forensic Science Australia and New Zealand, *A Guideline to Forensic Fundamentals: Identifying the Underpinning Science of Human Based Forensic Science Disciplines* (Report, 2016).

3 Donald Kennedy, 'Forensic Science: Oxymoron?' (2003) 302(5651) *Science* 1625; Laura Spinney, 'Science in Court: The Fine Print' (2010) 464 *Nature* 344.

4 Gary Edmond, 'Forensic Science Evidence and the Conditions for Rational (Jury) Evaluation' (2015) 39(1) *Melbourne University Law Review* 77.

5 NAS Report (n 2).

6 Sir Anthony Campbell, 'The Fingerprint Inquiry Report' (APS Group Scotland, 16 November 2011); Expert Working Group on Human Factors in Latent Print Analysis, 'Latent Print Examination and Human Factors: Improving the Practice Through a Systems Approach' (US Government Printing Office, 2012).

7 PCAST Report (n 2).

scientific underpinning. In particular, the report focused on the lack of evidence supporting the scientific validity and reliability of the forensic feature-comparison disciplines. The term 'forensic feature-comparison methods' is used in the PCAST Report in reference to

> the wide variety of methods that aim to determine whether an evidentiary sample (eg, from a crime scene) is or is not associated with a potential source sample (eg, from a suspect) based on the presence of similar patterns, impressions, features, or characteristics in the sample and the source.[8]

The report specified the need for empirical evidence showing that forensic disciplines conform to scientific standards for foundational validity — whether evidence is based on reliable principles and methods. The report also insisted that evidence is needed to show that forensic disciplines conform to scientific standards for applied validity — whether examiners reliably apply the principles and methods.[9] As the methods used in the feature-comparison disciplines depend upon human judgements to make evidentiary conclusions, evidence of foundational and applied validity needs to show that human judgement does indeed reliably produce accurate conclusions.

Examiners in the forensic feature-comparison disciplines are the 'instruments of analysis'[10] who deploy perceptual expertise in order to make judgements about forensic evidence.[11] Regardless of the object of examination (eg fingerprints, footwear, tyre tracks), forensic feature-comparison examiners engage in a cognitive process to evaluate the visual similarity between marks left at a crime scene and marks of known origins. Determining whether two marks came from the same or different sources is a subjective judgement based on an examiner's experience and expertise, and these judgements are presented to fact-finders as case evidence. For fact-finders to be able to rationally evaluate evidence, it is necessary to know how accurately forensic examiners make these conclusions.[12] A forensic examiner's conclusions are the output of human decision-making, which means that these conclusions contain an unavoidable potential for error.[13] In order for fact-finders to be able to evaluate the reliability and probative value of a forensic examiner's conclusions as case evidence, we

---

8 Ibid 23.

9 Ibid 43.

10 Itiel Dror and Simon Cole, 'The Vision in "Blind" Justice: Expert Perception, Judgment, and Visual Cognition in Forensic Pattern Recognition' (2010) 17(2) *Psychonomic Bulletin and Review* 161.

11 Matthew Thompson and Jason Tangen, 'The Nature of Fingerprint Expertise: Experts Can Do a Lot with a Little' (2014) 9(12) *PLOSONE* e114759.

12 Edmond (n 4).

13 Jason Tangen, 'Identification Personified' (2013) 45(3) *Australian Journal of Forensic Sciences* 315.

need empirical evidence demonstrating forensic examiner performance and limitations.[14]

In order to establish a scientific basis for forensic feature-matching disciplines, and provide fact-finders with information to apprehend the probative value of forensic testimony, empirical evidence of forensic examiners' abilities is needed. The publication of the NAS and PCAST Reports have led to research efforts to ground forensic science and expert testimony in empirical evidence,[15] but these efforts have been described as modest.[16] We argue that progress has been hindered by the absence of clear claims about what forensic examiners can do and how well they can do it. Without claims to knowledge or performance that lend themselves to empirical testing, the development of quantifiable measures of human performance in forensic pattern-matching cannot be satisfied.[17]

## III Empirical Testing Begins with a Claim

The epistemics of forensic science are beyond the scope of this article.[18] For our purposes, an empirical claim is one that can be verified by observation. More narrowly, an empirical claim can be supported, or not, by the results of experimentation. Once a claim is made, hypotheses can be generated, and predictions can be made of what will be observed if the claim is true. Such experiments provide the evidence required to empirically evaluate the truth of a claim. In order to be translated into empirical tests, claims need to be clear and specific, indicating the conditions of the claim and the criteria that must be satisfied in order to determine that the claim is true.

---

14 Jonathan Koehler, 'Proficiency Tests to Estimate Error Rates in the Forensic Sciences' (2013) 12 *Law, Probability and Risk* 89.

15 Much work has been done on fingerprints. See, eg, Thompson and Tangen (n 11); Matthew B Thompson, Jason M Tangen and Duncan J McCarthy, 'Human Matching Performance of Genuine Crime Scene Latent Fingerprints' (2013) 38 *Law and Human Behavior* 84; Jason M Tangen, Matthew B Thompson and Duncan J McCarthy, 'Identifying Fingerprint Expertise' (2011) 22 *Psychological Science* 995; Bradford T Ulery et al, 'Accuracy and Reliability of Forensic Latent Fingerprint Decisions' (2011) 108 *Proceedings of the National Academy of Sciences of the United States of America* 7733.

16 Jennifer Mnookin, 'The Uncertain Future of Forensic Science' (2018) 147 *Daedalus* 99.

17 Simon Cole, 'Who Will Regulate American Forensic Science?' (2018) 48 *Seton Hall Law Review* 563.

18 For a discussion of epistemological foundations of the forensic sciences, individualisation and the principle of uniqueness, see Michael Saks and David Faigenbaum, 'Failed Forensics: How Forensic Science Lost Its Way and How It Might Yet Find It' (2008) 4 *Annual Review of Law and Social Science* 142, 153–5; Jonathan Koehler and Michael Saks, 'Individualisation Claims in Forensic Science: Still Unwarranted' (2010) 75 *Brooklyn Law Review* 1187, 1191–8; Simon Cole, 'Forensics Without Uniqueness, Conclusions Without Individualisation: The New Epistemology of Forensic Identification' (2009) 8 *Law, Probability, and Risk* 233.

An illustrative example of how testable claims are articulated is presented in a challenge put forth in versions of the 'Paranormal Challenge' issued by sceptic communities.[19] They offer a financial prize to anyone who is able to demonstrate their proposed paranormal abilities in an empirical test.[20] In order to be able to design an empirical test of paranormal abilities, the challenge requires that interested participants provide a concise account of the claim. 'Then and only then', sceptics say, 'can we begin to design a test, one that is in accord with your claims.'[21] They ask three questions: (1) What exactly is your paranormal claim? (2) Under what conditions can you perform your claim? and (3) What success rate do you expect?[22]

Describing what a given claim is, the conditions necessary for the claim to be true, and the anticipated success rate allows one to generate a hypothesis that can be tested. A claim such as 'I can do X, under conditions Y, with a success rate of Z', lends itself well to empirical testing. We can hypothesise that if X is true, given Y we will observe X at a rate of ≥ Z. A more specific example of the translation of a claim into an empirical test involved a specialist group of Australian Passport Officers, who are relied upon to verify the identity of persons in order to administer valid passports.[23] It was hypothesised that the face-matching accuracy of passport officers would exceed that of student participants when tested in a laboratory setting where the ground truth is known.[24] These passport officers, however, were not any more accurate than student participants. The empirical evidence did not support the hypothesis and so did not lend support to the claim that people with specialist facial-comparison training and experience will outperform non-specialists at matching unfamiliar faces.

Identifying performance claims made by the forensic feature-comparison disciplines is the first step in designing empirical studies to test forensic examiners' abilities. All forensic feature-comparison disciplines ultimately rely on perceptual skill to evaluate evidence. These disciplines, however, work with different forms of evidence, operate using different methods and procedures, and are organised into independent professional bodies. Claims made by each forensic discipline may differ qualitatively — that is, claims of perceptual skill may not be transposable across different objects of examination — and so would require different approaches to measure performance and empirically test claims. In

---

19 The Skeptic's Dictionary, *Randi $1,000,000 Paranormal Challenge* (Web Page) <http://skepdic.com/randi.html>; Australian Skeptics, *The $100,000 Challenge* (Web Page) <https://www.skeptics.com.au/features/prize/>.

20 It is not our intent to invite comparison between forensic examiners and those who purport to have paranormal abilities, merely to provide a demonstration of what a directly testable claim looks like.

21 Cole (n 17).

22 Ibid.

23 David White et al, 'Passport Officers' Errors in Face Matching' (2014) 9(8) *PLOS ONE* e103510.

24 Ibid.

order to identify claims made by each forensic feature-comparison discipline, we will survey the professional literature published by key organisations that represent each discipline.

## IV Notes on Our Methods

In order to ascertain whether specific and testable claims are made by the forensic disciplines, we have surveyed the literature of professional society bodies representing the following pattern-matching disciplines: (1) Friction Ridge Examination (ie fingerprints); (2) Footwear and Tyre Tread Analysis; (3) Firearms and Toolmarks Analysis; and (4) Handwriting and Document Analysis. By focusing the survey of claims on literature from forensic science organisations, we will provide a view of performance claims made by disciplines as a whole rather than assigning the views of entire disciplines to statements made by individual experts in the pressurised circumstances of court proceedings. We attempt to find claims about forensic examiners' performance that are posited in such a way as to inform hypotheses that can be tested empirically. Such claims might appear in the form of propositions or statements about tasks that examiners are able to perform, including the conditions required for examiners to perform these tasks, and indicated standards of performance.

The survey is not intended to provide an exhaustive documentation of all claims that forensic examiners have made, but rather a representation of the information that is readily accessible to judges, lawyers, researchers and other interested parties. While it would be ideal to engage authorship of representatives from each forensic discipline, it is beyond the scope of this initial survey. Similarly, we do not include claims made by individual examiners in media interviews, or on websites for commercial services; nor do we include statements made by individual examiners presenting evidence in court. The literature surveyed is sourced from publications of key organisational bodies representing the forensic pattern-matching disciplines. Source material is limited to official publications that have been reviewed and endorsed by discipline representatives and made available on the websites of professional organisations. In Australia, the key organisational body is the National Institute of Forensic Science ('NIFS'), organised under the Australian and New Zealand Police Advisory Agency ('ANZPAA'). In Europe, forensic feature-matching disciplines are represented within the European Network of Forensic Science Institutes ('ENFSI'), which has undertaken responsibility for publishing standards of practice in the forensic sciences. In the United States, the forensic feature-matching disciplines are organised under the National Institute of Science and Technology ('NIST') Organisation for Scientific Area Committees ('OSAC'). OSAC subcommittees

> identify baseline documents and reference materials that best reflect the current state of the practice within their respective disciplines … that can help forensic scientists,

> judges, lawyers, researchers, other interested parties and the general public, to better understand the nature, scope, and foundations of the individual disciplines as they are currently practiced.[25]

We expect documents published by professional organisations to provide the best representation possible of current claims. We will offer the data and interpretation, which we do not claim to be definitive; others may interpret the data differently than we do. We do our best not to over interpret the data, and to leave room for disagreement. As such, and somewhat ironically, our claims about the forensic disciplines' claims are necessarily subjective and fuzzy. Nonetheless, the results provide a preliminary view of the current status of performance claims made by the forensic feature-matching disciplines.

## V Claims in Friction Ridge Examination (Fingerprints)

When a fingerprint is found at a crime scene, it is the job of a human fingerprint examiner to match the print to a known suspect or search for the print on a fingerprint database. The examiner will visually compare two prints, often on a computer screen, to decide whether the prints come from the same person or two different people. Fingerprint examination is described as a complex process consisting of visualisation (detection), imaging and individualisation, which itself consists of Analysis, Comparison, Evaluation and Verification ('ACE-V').[26]

At its core, fingerprint examination aims to determine whether two fingerprints came from the same or different sources. The standards outlined for fingerprint examination conclusions state:

> In reaching a conclusion, an examiner assesses the support of the observations for whether the two friction ridge impressions originated from the same source or from different sources. This document establishes the use of five conclusions: Source Exclusion, Support for Different Sources, Inconclusive/Lacking Support, Support for Same Source, and Source Identification.[27]

By 'assessing support of the observations', examiners conclude whether two fingerprints came from the same source or different sources. This is not, however, proffered as a claim of examiners' abilities. It does not state that examiners' conclusions are accurate to any degree. To falsify this statement would mean to demonstrate that examiners do not make conclusions by assessing the support of

---

[25] National Institute of Science and Technology Organisation of Scientific Area Committees for Forensic Science, *Annual Report: February 2016 – February 2017* (Report, 2017) 12 < https://www.nist.gov/sites/default/files/documents/2018/01/11/osac_annual_report_2017.pdf>.

[26] ENFSI, 'Best Practice Manual for Fingerprint Examination' (November 2015) 4 < http://enfsi.eu/wp-content/uploads/2016/09/6._fingerprint_examination_0.pdf>.

[27] Friction Ridge Subcommittee for NIST OSAC, 'Standard for Friction Ridge Examination Conclusions' (forthcoming) 4.

observations. However, given that the five conclusions stated are issued as case evidence and so are effectively presented as facts, an implicit claim can be inferred: that examiners make definitive source exclusion or identification conclusions accurately.

A description of fingerprint examiners' role is elaborated upon in a training document published by the Friction Ridge Subcommittee for the NIST OSAC:

> The examiner analyses impressions or other marks to detect relevant details, to compare these details to a reference exemplar, and to evaluate the probability of the observations under two competing propositions: (1) the two friction-ridge impressions originated from the same source, or (2) the two friction-ridge impressions originated from different sources. The examiner then forms opinions regarding (a) the weight or significance of the correspondence (or lack of correspondence) in the observed details of different marks or (b) which proposition is true.[28]

To state that examiners form either of two opinions — the 'significance of correspondence in details' or the truth of one of two exclusive propositions — is to make two distinct claims about examiners' ability. Opinions regarding the weight or significance of a degree of observed correspondence between features involve probabilistic estimates of the rate of correspondence occurring in the population. It is difficult to empirically test the performance accuracy of estimating probabilities. Given that it is not possible to determine population rarity in fingerprint features, we have no objective measure to compare performance to. However, probability estimates could be empirically tested for reliability between examiners — that is, the degree to which probability estimates made by different examiners converge. This would still require claims to stipulate the degree of reliability that needs to be observed to provide support for the claim. The statement that examiners determine which proposition is true entails the implicit claim that fingerprint examiners can identify whether two prints came from the same or different sources. Regardless of whether the truth of a proposition can, in practice, be known with absolute certainty, examiners' ability to accurately discern prints from the same and different sources can be tested empirically.

The European forensic body, ENFSI, makes no direct or indirect claims about examiners' ability to make source attributions. Rather, ENFSI states:

> The objective of the comparison is to determine the correspondences and/or dissimilarities between the features found during the analysis of the mark and the features found during the analysis of the reference print.[29]

---

28 Ibid.

29 ENFSI (n 26) 14.

This description refers to examiners' ability to identify similarities and differences between two fingerprints; unlike the aforementioned statements, it makes no claims of the ability of fingerprint examiners to link a print to a source, or to determine whether two prints came from the same or different sources. Additionally, it merely states that determining similarities and differences is the objective of comparison, not that this objective is achieved by examiners reliably or with a particular level of accuracy. Perception of similarities and differences in visual information is inherently subjective, and so it is not possible to measure performance on this task against a known truth. Examiners' detection of similarities and differences between prints could be tested for reliability between examiners; however, there is no such claim that performance on this task is reliable across examiners.

Some qualifications are issued in regard to examiners' ability to make conclusions as a result of fingerprint comparisons. Fingerprint examination is expressly purported to be not infallible and to not have a zero error rate.[30] Also, error rates are proposed to be conditioned on the quality of fingerprint samples, and this is encouraged to be represented in research into fingerprint examiners' performance.[31] Although these constraints provide some direction for empirical testing of fingerprint examiners' ability, they are not stated consistently across the fingerprint discipline, and are not formulated into direct and testable performance claims.

## VI Claims in Footwear and Tyre Tread Analysis

Footwear or tyre tread marks found at a crime scene might be cited as evidence of a particular shoe worn at, or tyres on a vehicle that moved through, that crime scene. Footwear and tyre tread examiners compare footwear and tyre tread marks recovered from a crime scene with marks of known origins, either by observing patterns side-by-side (physically, or in photographs on a computer screen), or by overlaying images.[32] Crime-scene marks may be imprints, residue on a hard surface (eg dust on a linoleum floor), or impressions (eg a pattern relief in mud). Patterns of known origin may be obtained from the footwear or vehicle tyres belonging to a suspect, or from databases of footwear and tyres.[33]

---

30 Ibid 5.

31 Friction Ridge Subcommittee for the NIST OSAC of Forensic Science, 'Response to PCAST Call for Additional References' (Report, 2016) 3.

32 A pattern copied onto a transparent film is overlaid on top of another to observe a fit in pattern features.

33 Databases are maintained by law enforcement agencies, or by footwear or tyre manufacturers.

No explicit claims of examiners' abilities were found to be made by professional bodies in the Footwear and Tyre Tread Discipline; however, an implicit claim is identified in statements made in Scope of Work and Examination Guide documents. This implicit claim refers to examiners' ability to 'include, identify, or eliminate a shoe or tire as the source of an impression',[34] based on visual comparison of footwear or tyre design.[35] This claim might seem readily subject to empirical tests, by asking examiners to visually compare pairs or sets of shoe prints and tyre marks and decide whether they are, might be, or are not from the same source. However, it is not clear what level of performance needs to be observed to provide support for the claim.

Implicit claims are also found in the conclusion scales that examiners use to present their comparison evidence to the courts. In the United States, results of footwear and tyre comparisons are typically reported using a conclusion scale that refers to the degree of association of design characteristics between a known and unknown footwear or tyre mark.[36] In Europe the conclusion scale developed by the ENFSI subcommittee is the common standard for footwear and tyre comparisons, which refers to the likelihood or probability of a particular shoe or tyre being the source of a print.[37] Although it is not expressly claimed in these scales that examiners consistently and accurately come to these conclusions, both of these conclusion scales suggest that footwear and tyre examiners can identify or eliminate a particular shoe or tyre as the source of an impression. Both scales also suggest that examiners can determine when there is not enough evidence to conclusively identify or eliminate a shoe or tyre as the source of an impression ('inconclusive' and 'lacks sufficient details'). The remaining intermediary conclusion options pertain to different types of information that examiners can discern from footwear or tyre evidence: degree of similarity in features, or the probability of a shoe or tyre being the source of a print. Examiners' conclusions

---

34 Footwear and Tire Subcommittee of the OSAC for Forensic Science (Scientific Working Group for Shoeprint and Tire Tread Evidence), *Scope of Work Relating to Forensic Footwear and/or Tire Tread Examiners*, (March, 2005) [2.1] <http://treadforensics.com/images/swgtread/standards/current/swgtread_06_scope_of_work_200503.pdf.>

35 Footwear and Tire Subcommittee of the OSAC for Forensic Science (Scientific Working Group for Shoeprint and Tire Tread Evidence), *Guide for the Examination of Footwear and Tire Impression Evidence* (March, 2006) [6.7] <http://treadforensics.com/images/swgtread/standards/current/swgtread_08_examination_200603.pdf>.

36 OSAC Footwear and Tire Subcommittee, *Response to PCAST Request for Information* (December 2015). The SWGTREAD conclusion scale is as follows : (1) lacks sufficient detail, (2) exclusion, (3) limited associated of class characteristics, (4) association of class characteristics, (5) high degree of association, and (6) identification.

37 Ibid. The ENFSI conclusion scale is as follows: (1) elimination, (2) likely not the source of the impression, (3) inconclusive, (4) probably the source of the impression, (5) very probably the source of the impression, and (6) identification.

about degrees of similarity or source probability could be tested for reliability between examiners, rather than accuracy. This would require claims to indicate the level of reliability to be observed to provide empirical support for the claim.

Additionally, a position document issued in response to the NAS Report states that examiners can use footwear and tyre tread evidence in determining: number of perpetrators; presence at a crime scene; path of travel through the scene; relevant footwear or tyre information; and corroborating or refuting statements from witnesses and suspects.[38] This statement refers to an ability to determine activity-level information from footwear and tyre impressions, making additional inferences about events that occurred in relation to a crime beyond linking an impression to a source. No claim is made as to examiners' ability to correctly derive this information.

Although the primary function of footwear and tyre tread examiners is to express an opinion as to whether two footwear or tyre prints or impressions came from the same or different sources, this is not expressed as a direct and testable claim. No standards of accuracy are indicated for examiners' decisions. Furthermore, print and impression evidence can appear in varying substrates and degrees of quality, and no claims are made regarding abilities being contingent on the nature of evidence. Without specific claims indicating conditions of examiners' abilities, the implicit claim appears to be that the ability of examiners to detect whether two impressions came from the same or different sources is generalisable across all forms of footwear and tyre evidence, and is not liable to vary with the quality of impression or print evidence.

## VII Claims in Firearm and Toolmark Analysis

Tools may be used in the commission of crimes — for example, in forcing entry. When a hard object makes contact with a softer object, it will impart depressions, marks and scratches constituting a toolmark. Toolmarks found at a crime scene are either cut out of the surface they are found in, or cast in silicone and transferred to a laboratory where they are analysed by toolmark examiners. If a tool is recovered from a crime scene or from a suspect's possession, examiners will create test marks using this tool to compare these marks with toolmarks detected at the crime scene. Comparisons involve the use of a comparison microscope, which provides examiners with a side-by-side view of miniscule

---

[38] Scientific Working Group for Shoeprint and Tire Tread Evidence, 'Current Status of the Forensic Footwear and Tire Tread Examination Discipline' (Document 1, March 2009) <http://treadforensics.com/images/documents/responses/nas/nas_response_current_status.pdf>.

striations comprising marks.[39] Firearms identification is recognised as a subset of the toolmark identification forensic science discipline.[40] As with other suspect tools, a suspect gun is test-fired in order to recover ammunition components (ie bullets and cartridge cases), which serve as reference samples wherein the microscopic marks can be compared with those on ammunition components recovered from a crime scene.[41]

The Firearm and Toolmark Discipline makes the claim that trained examiners can reliably differentiate and associate marks made by the same or different tools by visually comparing microscopic marks.[42] This ability enables examiners to determine whether a particular tool is the source of a mark. The claim that examiners can identify a tool as the source of a mark is implied in the conclusion scales used by examiners in presenting the outcome of their comparisons as evidence;[43] however, the conditions required for examiners to perform this task is unclear.

Although the conditions required for examiners to perform their claimed abilities are not explicitly stipulated, the Association of Firearm and Tool Mark

---

39 National Institute of Forensic Science Organisation of Scientific Area Committees, *Firearms and Toolmarks Subcommittee* (Web Page) <https://www.nist.gov/topics/forensic-science/firearms-and-toolmarks-subcommittee>.

40 AFTE Committee for the Advancement of the Science of Firearm and Toolmark Identification, 'The Response of the AFTE to the NAS 2008 Report Assessing the Feasibility, Accuracy, and Technical Capability of a National Ballistics Database' (August 2008) 40(3) *AFTE Journal* 234.

41 Stephen Bunch et al, 'Is a Match Really a Match? A Primer on the Procedures and Validity of Firearm and Toolmark Identification' (2009) 11(3) *Forensic Science Communications* <https://www2.fbi.gov/hq/lab/fsc/backissu/july2009/review/2009_07_review01.htm>.

42 This claim is stated explicitly in a publication on the procedures and validity of firearm and toolmark identification. See AFTE Committee (n 40) 237: 'the proposition that trained firearm-toolmark examiners can distinguish between marks made by either the same or different firearms or tools'. See also Firearm and Toolmark Subcommittee of the National Institute of Standards and Technology (NIST) Organisation of Scientific Area Committees (OSAC) for Forensic Science, 'Response to the President's Council of Advisors on Science and Technology (PCAST)' (23 Dec 2015) <https://www.nist.gov/sites/default/files/documents/2016/12/14/osac_firearms_toolmarks_subcommittees_response_to_the_pcast_request_for_info.pdf>: 'In firearms and toolmark identification, performance testing … is determined by a series of different tests and experiments. First, the overall reliability of a trained examiner to correctly differentiate and associate items based on the comparison of microscopic toolmarks.'

43 Firearms and Toolmarks Committee of the OSAC for Forensic Science, 'Standard Scale of Source Conclusions and Criteria for Firearm and Toolmark Examinations' (Draft Document) [4, 4.2.5.1] <https://www.nist.gov/sites/default/files/documents/2019/04/12/fatm_roc_and_criteria_standard_asb_mar2019.pdf>: '(1) Exclusion, (2) Insufficient support for exclusion, (3) Insufficient support for either exclusion or identification, (4) Insufficient support for identification, (5) Identification. An Identification conclusion is based on an examiner's determination that all discernible class and individual characteristics agree such that *the extent of agreement exceeds that which has been demonstrated by toolmarks made by different tools and is consistent with the agreement demonstrated by toolmarks known to have been made by the same tool*' (emphasis added).

Examiners ('AFTE') states that examiners are not limited to or specialised in any particular kind of toolmark, and can make accurate source conclusions about toolmarks regardless of the type of tool that made the mark and the surface the toolmark is made in.[44] Several sources refer to the level of accuracy at which examiners perform, stating that they 'rarely, if ever' make errors when performing this claim.[45] However, this assertion does not lend itself to empirical testing; it is not clear how 'rarely' should be quantified, that is, what level of accuracy needs to be observed to support claims of toolmark examiners' ability to discriminate toolmarks made by the same and different tools.

## VIII Claims in Handwriting and Document Analysis

Handwriting and document examiners provide evidence to the courts concerning the authenticity of signatures, written documents, machine generated documents, ink comparisons and recovery of information from damaged documents. The Handwriting and Document Discipline proffers that forensic document examiners are able to determine the source of a piece of handwriting and, although this is not a direct claim, it is suggested that this is so across multiple professional organisations.[46] It is unclear whether examiners require a

---

44 AFTE Committee (n 40) 238: 'Toolmark examiners are trained to examine the marks left by tools on any variety of surfaces in an attempt to associate a toolmark to a particular tool that produced the mark … The approach to examining toolmarks of any and all types is the same whether the toolmarks were generated by a screwdriver tip, a pry bar or the bore or breechface of a firearm.'

45 In a Position Statement in response to the NAS Report (n 40) 238, the Association for Firearm and Toolmark Examination states: 'Toolmarks imparted to objects by different tools (known non-matches) will rarely if ever display agreement sufficient to lead a qualified examiner to conclude the objects were marked by the same tool. That is, a qualified examiner will rarely if ever commit a false positive error (misidentification)'. See also SWGGUN Foundational Report, republished in Firearm and Toolmark Examination Subcommittee for the NIST OSAC (Scientific Working Group for Firearms and Toolmarks), 'The Foundations of Firearm and Toolmark Identification' (2013) 5 <https://www.nist.gov/sites/default/files/documents/2016/11/28/swggun_foundational_report.pdf>: 'despite the subjective nature of the analysis, competent examiners will rarely, if ever, commit false identifications or false eliminations'.

46 The American Society of Questioned Document Examiners ('ASQDE'), *Frequently Asked Questions* (Web Page) <http://www.asqde.org/about/faq.html>: 'For most forensic document examiners, the most common examination is the comparison of handwriting and signatures to determine whether someone did or did not write them'. See also ENFSI, 'Best Practice Manual for the Forensic Examination of Handwriting' (Version 2, June 2018) 3 < http://enfsi.eu/wp-content/uploads/2017/06/Best-Practice-Manual-for-the-Forensic-Examination-of-Handwriting-Version-02.pdf>: 'the scientific examination and comparison of handwritten documents to determine whether or not two or more pieces of handwriting have been completed by one individual. This includes authentication of one or more questioned signatures by comparison with a set of known signatures'; Southwestern Association of Forensic Document Examiners ('SWAFDE'), *Frequently*

set of exemplars or a single exemplar in order to make a judgement about the source of handwriting. Specifying this requirement in empirical claims is necessary to devise empirical tests of examiners' ability to discern the source of handwriting.

The ability of examiners to link a piece of handwriting to a source is qualified in several statements, indicating some conditional constraints on examiners' ability to link a piece of writing to a source. These include: 'the questioned or known writing may be too limited, or the questioned writing may be disguised to such a degree as to prevent identification';[47] and that examiners are only able to link a piece of handwriting to a source if there is a sufficient sample of handwriting from the known source, and that this known sample is representative of the elements in the questioned handwriting (eg there are common elements, such as letters).[48] Stating that examiners may not be able to identify a piece of handwriting if it is disguised to a degree that prevents identification seems somewhat incompatible with source attribution claims. It is unclear what it means for a piece of handwriting to be disguised. Handwriting may be disguised in an attempt at forgery, in which case this claim can be interpreted as examiners' being unable to identify the source of handwriting if it has been forged so as to resemble the source.

In addition to handwriting examination, forensic document examiners conduct investigations concerning printed documents. According to one description of forensic document examination, examiners have the capacity to 'identify or eliminate the source of machine-produced documents, typewriting, or other impression marks, or relative evidence, and preserve and/or restore legibility'.[49] This description makes no direct claims of ability, but suggests that forensic document examiners can link a document to the source (ie machine) that produced it. Other professional organisations refer to the ability of examiners to

---

*Asked Questions* (Web Page) <http://www.swafde.org/faq/>: 'A Forensic Document Examiner examines documents to determine authenticity and/or to discover who wrote them … The most common examination is the comparison of handwriting and/or signatures'; Scientific Working Group for Handwriting and Document Analysis ('SWGDOC'), 'Standard for Scope of Work of Forensic Document Examiners' (2015) <https://www.swgdoc.org/documents/SWGDOC%20Standard%20for%20Scope%20of%20Work%20of%20Forensic%20Document%20Examiners.pdf>: 'The forensic document examiner conducts scientific examinations, comparisons, and analyses of documents in order to: (1) establish genuineness or nongenuineness, or to reveal alterations, additions, or deletions, (2) identify or eliminate persons as the source of handwriting'.

47 SWAFDE (n 46).

48 Ibid: 'The general rule is that hand-printing can only be compared with hand-printing, and handwriting must be compared with handwriting. Also, there must be similar text in both documents to be compared. For example, 'Jack' cannot be compared with 'Bob' because there are no common letters.'

49 SWGDOC (n 46).

be able to preserve, restore and ultimately decipher evidence, including documents subject to myriad kinds of damage, including deliberate alterations, erasure and charring.[50] These different conditions would constitute independent claims of examiners' ability to decipher evidence.

We found no specific claims regarding the accuracy with which handwriting and document examiners are proposed to determine the source of handwriting or documents. Some indication as to the level of accuracy that should be observed to satisfy these claims is needed to be able to devise hypotheses that can be tested to derive empirical evidence of handwriting and document examiners' abilities.

## IX Claims Across Disciplines

There were implicit claims made by each discipline referring to the particular tasks that examiners perform and the judgements they make, and in some instances the conditions required for examiners to perform claims were stipulated. None of the disciplines quantified the lower level of performance that is expected to be demonstrated by examiners in order to satisfy an agreed-upon standard, expectation or acceptable level. This state of affairs may be an impediment to establishing the foundational validity of the forensic disciplines, because we do not know what empirical evidence is needed to support performance claims.

Through the lens of cognitive psychology, all forensic feature-comparison disciplines engage in similar perceptual processes to decide whether two objects came from the same or different sources. This notion is reflected in statements of examiners' abilities made across the forensic feature-matching disciplines, which consistently allude to the ability to distinguish objects (eg marks and prints) that have come from the same or different sources. Source-attribution claims across the forensic feature-matching disciplines provide a starting point for the disciplines to formulate empirically testable claims. But there appear to be

---

[50] ASQDE (n 46): 'Other types of examinations include the examination typewriting, computer printed documents, photocopies, decipherment of altered, obliterated and charred documents, the examination of inks and paper, decipherment of erased entries and indented writings, detection of counterfeit currency, and the examination of commercially printed matter'; SWAFDE (n 46): 'Other examinations include examination of typewritten or machine-generated documents; detection of alterations; decipherment of obliterated and indented writing; examination of watermarks, rubber stamps, and other impressions; and ink differentiation'. Collectively, these statements refer to document examiners' ability to (a) identify or exclude the machine source of a document, (b) detect alterations made to documents, (c) decipher original information in documents that have been altered or damaged, including charred documents, (d) differentiate between different inks used in documents, and (e) detect counterfeit documents, including currency. Once again, there are no explicit claims about examiners' performance of these tasks.

inconsistencies that need to be addressed, most notably the conditions necessary for the apprehension of intra-individual variation in objects when determining whether two objects came from the same or different sources. The Fingerprint, Footwear and Tyre Tread, and Firearms and Toolmarks Disciplines seem to claim to be able to identify two objects as having come from the same or different sources, implying an appreciation of the distinction between variation that occurs between marks made by the same source, and variation that occurs between marks made by different sources. In contrast, the Handwriting and Document Discipline seems to claim that source-attributions can only be made given a sufficient sample of known exemplars in order to account for the variation that occurs in objects produced by a single source.[51]

## X Conclusion and a Formula for the Future

In this article, we set out to identify and evaluate the performance claims made by the forensic feature-matching disciplines. We surveyed the professional literature published by representative organisations for Friction Ridge Examination, Footwear and Tyre Tread Analysis, Firearm and Toolmark Analysis, and Handwriting and Document Analysis. We found that many claims did not readily lend themselves to empirical testing against our proposed general framework of: 'I can do X, under conditions Y, with a success rate of Z.' Implicit claims were identified in descriptions of forensic examiners' scope of work, discipline representatives' responses to investigative reports, and in the conclusion scales used by examiners. All feature-matching disciplines made implicit claims pertaining to examiners' ability to link marks to a source by visual comparison of questioned marks and marks of known origin. Although the implicit claims we have identified provide a starting point for developing empirical tests of the forensic disciplines, these claims are incomplete.

Because the forensic feature-matching disciplines do not readily articulate empirically testable claims of examiners' ability, we do not know what kind of tests and evidence are needed to establish empirical evidence of examiners' abilities. Empirical testing of examiners' performance is necessary to obtain information about the accuracy and reliability of examiners' judgements, in order to establish foundational validity for the forensic disciplines. As we have argued, this information is necessary to establish indicative rates of performance and an appreciation of limitations to enable fact-finders to rationally evaluate forensic evidence. Current claims appear to be imprecise and may not lend themselves to empirical testing. We propose the following formula in order to make progress:

---

[51] The Firearm and Toolmark Discipline does state that, given a suspect tool, multiple test marks are made to provide a set of marks for the basis of comparison, but this is not included in their claim that examiners can distinguish between marks made by the same or different tools.

1. Where implicit claims are made, but they are not specific or directly testable claims of examiners' performance abilities, it would be prudent to make these abilities explicit ('I can do X'). This is a necessary first step to creating hypotheses and designing empirical tests to provide empirical evidence of abilities.
2. If a claim has been made, but it is not testable, amend the claim to make it testable. Given what examiners are proposed to be able to do, identify the conditions required for them to be able to do it ('I can do X, under conditions Y'). This way, the proposed conditions, or factors affecting performance, can be accounted for in empirical studies.
3. Identify a level of performance that examiners are expected to achieve so that empirical evidence demonstrating this level of accuracy can be said to provide support for the claim ('I can do X, under conditions Y, with a success rate of Z'). The level could be, for example, a comparison to statistical chance or to laypersons, or to a standard based on professional consensus, policy or legislation.

Once testable claims are made, we can begin to make progress. Hypotheses can be formulated towards providing support for a claim. Hypotheses will determine how performance of the claim is enacted and measured in order to produce empirical evidence that the claim is supported. Once there are hypotheses, they can be empirically tested (over, and over again) in order to amass an empirical evidence base on which forensic examiners' conclusions can rest. The evidence of abilities and limitations can then be shared with the scientific and legal communities. Once a performance baseline has been established, the factors that limit and augment examiners' performance can be determined, along with the optimal way to train and recruit forensic examiners. With a well-established base of empirical evidence, the best way for examiners to communicate opinion evidence to the courts can then be studied.

We acknowledge that attaining these goals will take time and commitment, but envisage that following our proposed (or a similar) formula would not only enable the forensic disciplines to surpass the minimum standards for scientific validity, but also engender continuous improvement to further enhance the contribution of forensic science to the legal system.