Advancing a paradigm shift in evaluation of forensic evidence – Part I: The rise of forensic data science

Authors and affiliations:

Geoffrey Stewart Morrison [1,2,*]

[1] Forensic Data Science Laboratory, Aston University, Birmingham, UK

[2] Forensic Evaluation Ltd, Birmingham, UK

*Corresponding author: G.S. Morrison, e-mail: geoff-morrison@forensic-evaluation.net

ORCID:

Geoffrey Stewart Morrison 0000-0001-8608-8207

Disclaimer:

All opinions expressed in the present paper are those of the author, and, unless explicitly stated otherwise, should not be construed as representing the policies or positions of any organizations with which the author is associated.

Declaration of competing interest:

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements:

This research was supported by Research England's Expanding Excellence in England Fund as part of funding for the Aston Institute for Forensic Linguistics 2019–2022.

I dedicate this paper in memory of Prof Terrance M Nearey, my former PhD supervisor, who taught me how to be a scientist. Terry passed away on December 18, 2021.

Advancing a paradigm shift in evaluation of forensic evidence – Part I: The rise of forensic data science

Abstract

Widespread practice across the majority of branches of forensic science uses analytical methods based on human perception and interpretive methods based on subjective judgement. These methods are non-transparent and are susceptible to cognitive bias, interpretation is often logically flawed, and forensic-evaluation systems are often not empirically validated. We describe a paradigm shift in which existing methods are replaced by methods based on relevant data, quantitative measurements, and statistical models; methods that are transparent and reproducible, are intrinsically resistant to cognitive bias, use the logically correct framework for interpretation of evidence (the likelihood-ratio framework), and are empirically validated under casework conditions.

Keywords

forensic science; forensic data science; likelihood ratio; paradigm shift; validation

1 Introduction

The present paper is published in two parts. Part I describes an ongoing paradigm shift in evaluation of forensic evidence. It describes the current state of affairs (*staus quo*), the new paradigm (*quo vadis?*), obstacles to the advancement of the paradigm shift (*impedimenta*), and a strategy to advance the paradigm shift (*via progredi*). It also describes approaches to dealing with problems presented by the new paradigm, in particular, how to calculate likelihood ratios for forensic data and how to validate forensic-evaluation systems that output likelihood ratios.

Part of our strategy involves collaboration between forensic-data-science researchers and researchers and practitioners in particular branches of forensic science who want to adopt the new paradigm. In any branch of forensic science, we expect the number of practitioners who initially want to adopt the new paradigm to be a very small, but it will be more productive to work with a small minority on developing practical solutions than to try to convince the majority without providing practical solutions. Once the practical solutions are being used by the small minority, use of the new paradigm has the potential to spread.

Part II (Basu et al., 2022) will provide an example of the application of part of our strategy. It will describe building a relevant database and developing and validating statistical models for forensic comparison of fired cartridge cases, a common tasks in forensic firearm examination, a branch of forensic science in which the new paradigm has so far made almost no progress. It focuses in particular on testing different functional-data-analysis-based methods for feature extraction from 3D images of fired cartridge cases. This will be a first step in attempting to advance the paradigm shift in forensic firearm examination, and is intended to be an example that can be copied in other branches of forensic science.

2 A paradigm shift in evaluation of forensic evidence

2.1 Status quo

Curran (2013):

> Is forensic science the last bastion of resistance against statistics?

UK House of Lords Science and Technology Committee (HoL, 2019):

> In regard to pattern comparison methods, ... “the comparison of fingerprints, toolmarks, footwear, tire marks and ballistics” [are] “spot-the-difference”

> techniques in which "there is little, if any, robust science involved in the analytical or comparative processes used and as a consequence there have been questions raised around the reproducibility, repeatability, accuracy and error rates of such analysis." (§155)

In forensic science, the process of *evaluation of strength of evidence* consists of: *analysis*, i.e., extraction of information from items of interest (the evidence);[1] and *interpretation*, i.e., drawing inferences with respect to the meaning of the information extracted by the analysis. Items of interest may be, for example: a fingermark of questioned source recovered from a crime scene and a fingerprint collected from a known individual; a recording of a speaker of questioned identity on an intercepted telephone call and a recording of a police interview with a speaker of known identity; a fired cartridge case recovered from a crime scene and cartridge cases fired in a forensic laboratory from a gun found in the possession of a suspected shooter. Forensic practitioners conduct evaluations in order to assist legal-decision makers to make decisions with respect to questions of legal concern such as: Do the fingermark and fingerprint originate from the same finger? Is the speaker of questioned identity on the intercepted recording the same as the speaker of known identity? Was the cartridge case recovered from the crime scene fired from the suspect's gun?[2]

Currently, across the majority of branches of forensic science, widespread practice is that analysis is conducted using *human perception*, and interpretation is conducted using *subjective judgement*. Even in branches of forensic science in which analysis is conducted using instrumental measurement, interpretation is commonly based on subjective judgement, e.g., by eyeballing graphical representations of the measured

[1] In the forensic-inference-and-statistics literature "evidence" is the term commonly used to refer to the items of interest (i.e., the input to the analysis) or to the information output by the analysis (i.e., the input to the interpretation). Usage is somewhat fluid, but, either way, this is evidence from the perspective of the forensic practitioner. From the perspective of the court, evidence is the oral testimony or written submission of the forensic practitioner. In wider forensic-science literature, the term "trace" is often used to refer to items of interest.

[2] The present paper is framed in the context of source-level comparison.

values. Human-perception-based analysis methods and subjective-judgement-based interpretation methods are non-transparent and susceptible to cognitive bias. Across the majority of branches of forensic science, even branches of forensic science in which interpretation is conducted using statistical models, interpretation of evidence is often logically flawed, and forensic-evaluation systems (the end-to-end combination of analysis and interpretation methods) are often not empirically validated or not adequately empirically validated.[3]

2.2 Quo vadis?

2.2.1 Introduction

Saks & Koehler (2005):

> we envision a paradigm shift in the traditional forensic identification sciences in which untested assumptions and semi-informed guesswork are replaced by a sound scientific foundation and justifiable protocols. Although obstacles exist both inside and outside forensic science, the time is ripe for the traditional forensic sciences to replace antiquated assumptions of uniqueness and perfection with a more defensible empirical and probabilistic foundation. (p. 895)

US President's Council of Advisors on Science and Technology (PCAST, 2016):

> neither experience, nor judgment, nor good professional practice ... can substitute for actual evidence of foundational validity and reliability. The frequency with which a particular pattern or set of features will be observed in different samples, which is an essential element in drawing conclusions, is not a matter of "judgment." It is an empirical matter for which only empirical evidence is relevant. (p. 6)

[3] Claims made in this introductory paragraph are supported by details and references provided later in §2.2 and in §2.3.

> Objective methods are, in general, preferable to subjective methods. Analyses that depend on human judgment (rather than a quantitative measure ...) are obviously more susceptible to human error, bias, and performance variability across examiners. In contrast, objective, quantified methods tend to yield greater accuracy, repeatability and reliability, including reducing variation in results among examiners. Subjective methods can evolve into or be replaced by objective methods. (p. 47)

A *paradigm shift* in evaluation of forensic evidence is ongoing in which methods based on human perception and subjective judgement are being replaced by methods based on *relevant data*, *quantitative measurements*, and *statistical models*; methods that:

1. are *transparent and reproducible*;
2. are *intrinsically resistant to cognitive bias*;
3. use the *logically correct framework for interpretation of evidence* (the *likelihood-ratio framework*); and
4. are *empirically validated under casework conditions*.

We address each of these elements in the following four subsections (§2.2.2–§2.2.5).

2.2.2 Transparency and reproducibility

Methods dependent on human perception and subjective judgement are intrinsically non-transparent and therefore not reproducible by others. Human introspection is often mistaken, hence a forensic practitioner's explanation of how they reached their conclusion may not reflect how they actually reached that conclusion (Edmond et al., 2017). In contrast, procedures based on data, quantitative measurement, and statistical models are transparent and reproducible: measurement (feature-extraction) and statistical-modelling methods can be described in detail, and data and software tools can potentially be shared with others.

2.2.3 Cognitive bias

There has been a great deal of concern about cognitive bias in forensic science (National Research Council, 2009; Expert Working Group on Human Factors in Latent Print Analysis, EWG, 2012; Found, 2015; Stoel et al., 2015; PCAST, 2016; Edmond et al., 2017; Cooper & Meterko, 2019). Cognitive bias is subconscious bias, it cannot be controlled by strength of will. Forensic practitioners are susceptible to cognitive bias when making perceptual observations: their belief in the probability that a hypothesis is true can affect their analysis of the evidence and therefore the information that feeds into their interpretation. Forensic practitioners are susceptible to cognitive bias when they are making subjective judgements and are exposed to information that could influence their belief in the probability that a hypothesis is true but that would not logically affect the probability of obtaining the evidence conditional on whether the hypothesis were true. Some potentially biasing information is task-irrelevant and should be withheld from practitioners, but some potentially biasing information is task-relevant and practitioners employing human-perception and subjective-judgement methods will need to be exposed to it at some point in the evaluation process, e.g., practitioners who visually compare known-source fingerprints and questioned-source fingermarks must be exposed to both, but exposure to a higher-quality print may bias their analysis of ambiguous details in a lower-quality mark. Systems in which the strength-of-evidence conclusion is directly the result of subjective judgement are particularly susceptible to cognitive bias.

Systems based on quantitative measurements and statistical models require subjective judgements on decisions such as whether the data used for training the system and the data used for validating the system are sufficiently representative of the relevant population for the case and sufficiently reflective of the conditions of the items of interest in the case so that the output of the system will be a meaningful answer to the question posed in the case and so that the results of the validation will provide a meaningful indication of the performance of the systems under the conditions of the

case. These decisions, however, are made at the beginning of the process before the practitioner has analyzed the items of interest, hence the practitioner cannot know what effect these decisions will have on the strength-of-evidence conclusion. The remainder of the evaluation process is automated, hence not susceptible to cognitive bias.

2.2.4 Likelihood-ratio framework

In current practice, interpretation of evidence is often logically flawed, e.g., it is based on the uniqueness or the individualization fallacy (Saks & Koehler, 2008; Cole, 2009, 2014), and conclusions are often expressed categorically, e.g., “identification”, “inconclusive”, “exclusion” (i.e., posterior probability of 1 or 0 with respect to the same-source hypothesis, with “inconclusive” meaning no conclusion rather than an intermediate probability), or using some form of uncalibrated verbal posterior-probability scale, e.g., “identification”, “probable identification”, “inconclusive”, “probable exclusion”, “exclusion”. Jackson (2009) and Kaye (2015) review these and other commonly used but logically flawed conclusions.

In contrast, the likelihood-ratio framework is advocated as the logically correct framework for evaluation of evidence by the vast majority of experts in forensic inference and statistics (including: Aitken et al., 2011; Morrison et al., 2017; Morrison, Enzinger, et al., 2021; with 31, 19 and 20 authors and supporters respectively), and by key organizations including: Association of Forensic Science Providers of the United Kingdom and of the Republic of Ireland (2009); Royal Statistical Society (Aitken et al., 2010); European Network of Forensic Science Institutes (Willis et al., 2015); National Institute of Forensic Science of the Australia New Zealand Policing Advisory Agency (Ballantyne et al., 2017); American Statistical Association (Kafadar et al., 2019); Forensic Science Regulator for England & Wales (FSR, 2021).

The likelihood-ratio framework requires assessment of the probability of obtaining the evidence if one hypothesis were true versus the probability of obtaining the evidence if an alternative hypothesis were true. The two hypotheses must be mutually exclusive.

One hypothesis should represent the position of the prosecution in the case, and the other the position of the defence, e.g., the fingermark of questioned origin was deposited by a finger of a particular known individual, versus the fingermark of questioned origin was deposited by a finger of some other individual selected at random from the relevant population. In this source-level fingerprint-fingermark comparison example, the numerator of the likelihood ratio quantifies the *similarity* between the mark and the print, and the denominator quantifies the *typicality* of the mark with respect to the relevant population. For continuously-valued data, likelihood ratios can be calculated as the ratio of two probability-density functions evaluated at the value of the evidence. In the forensic inference and statistics literature "likelihood ratio" is commonly used as a cover term for both likelihood ratios based only on sample data and for Bayes' factors based on sample data and prior distributions for model parameters. The present discussion is intended to be neutral with respect to likelihoodist or Bayesian approaches.

2.2.5 Empirical validation

Empirical validation under conditions reflecting those of the case to which a forensic-evaluation system is to be applied is the only way to know how well that system performs under the conditions of the case. Protocols for validating systems that output likelihood ratios have been developed, including metrics and graphics appropriate for representing the results of such validations (Morrison, 2011; Meuwly et al., 2017; Ramos et al., 2020; Morrison, Enzinger, et al., 2021). Much of the latter development has been conducted in the context of forensic voice comparison, but the results are applicable across forensic science in general. The need for validation under casework conditions has been emphasized by FSR (2020b), and by PCAST (2016):

> Without appropriate estimates of accuracy, an examiner's statement that two samples are similar—or even indistinguishable—is scientifically meaningless: it has no probative value, and considerable potential for prejudicial impact.

> Nothing—not training, personal experience nor professional practices—can substitute for adequate empirical demonstration of accuracy. (p. 46)

Despite this, practitioners in multiple branches of forensic science often claim that training and experience provide sufficient warrant for their conclusions (see Mnookin et al., 2011; Risinger, 2013; PCAST, 2016; Morrison & Thompson, 2017), deny or obfuscate about the need for validation (see Cole, 2006; Morrison, 2014; PCAST, 2016; Koehler, 2017; Morrison et al., 2018), or propose lax validation protocols that do not require demonstration of performance under casework conditions (see Morrison, Neumann, et al., 2020, 2021).

2.2.6 A Kuhnian paradigm shift

The idea that evaluation of forensic evidence is undergoing a paradigm shift is not new. The most famous article heralding a paradigm shift is Saks & Koehler (2005). Allowing for differences in wording and level of detail, and allowing for difference due to the passage of time, we believe that Saks & Koehler (2005) and the present paper describe the same paradigm shift. In contrast to Saks & Koehler's (2005) statement that they intended "paradigm shift" as a metaphor, however, we view the paradigm shift in evaluation of forensic evidence as a true Kuhnian paradigm shift (Kuhn, 1962) in the sense that it requires rejection of existing methods and the ways of thinking that underpin them, and rejection of the idea that progress can be made by incremental improvements to existing methods. Instead, it requires the wholesale adoption of an entire constellation of new methods and new ways of thinking.[4]

That a paradigm shift requires the wholesale adoption of an entire constellation of new methods and new ways of thinking remains the case irrespective of whether one considers the shift to be from one paradigm to another or to be from a pre-paradigm to

[4] The only aspects in which we think the current paradigm shift in evaluation of forensic evidence deviates from Kuhn's (1962) description of paradigm shifts relate to forensic science being an applied science which is not isolated from societal pressures.

a paradigm period of science. As suggested in Saks & Koehler (2005), a pre-paradigm period would seem to be a more accurate description of the *status quo*, with multiple traditions of evaluation of evidence used both within individual branches of forensic science and across different branches of forensic science, and hence there being no established widely-accepted overarching paradigm in use.

Some authors have used the term "paradigm shift" in relation to a single element or a subset of the elements of the paradigm shift as we have outlined it above, but we believe that all of these elements are required as part of the constellation. This may be viewed as a radical stance, and it faces resistance, but over the last decade and a half we have had substantial success in contributing to advancing this paradigm shift in forensic voice comparison.[5]

Over many years, multiple people have suggested to us that we give the new paradigm a name, rather than simply referring to it as the new paradigm. We hereby suggest that the new paradigm can be referred to as *forensic data science*, on the condition that the definition of forensic data science is understood to encompass the new paradigm as we have described it above.

2.3 Impedimenta

2.3.1 Introduction

The paradigm shift in evaluation of forensic evidence is ongoing, but progress is slow or stalling for multiple reasons, including the following:

1. The new paradigm has only been adopted in a few branches of forensic sciences, and only by a minority of researchers and practitioners.

2. Only some elements of the new paradigm have been adopted as part of

[5] "we" here refers to Morrison and colleagues in general.

incremental change.

3. There is misunderstanding of the new paradigm and resistance to its adoption.

4. Research is often not informed by practice and has no impact on practice.

5. It is difficult to obtain funding for evidential-forensic-science research.

6. There are genuine practical impediments to implementing the new paradigm.

We discuss each of these impediments in the following six subsections (§2.3.2–§2.3.7).

2.3.2 The new paradigm has only been adopted in a few branches of forensic sciences, and only by a minority of researchers and practitioners

The new paradigm has only been adopted in a few branches of forensic sciences, and only by a minority of researchers and practitioners. In the 1990s, the new paradigm was widely adopted for **forensic evaluation of DNA** (Foreman et al., 2003). Although the volume and importance of casework in this branch of forensic science makes it influential, single-source DNA profiles are invariant and discrete, and therefore have a very different structure from the continuously-valued data with within-source variability that results from analyses in most other branches of forensic science. The situation is more complex for low-template DNA and for DNA mixtures, but there is still a difference in data structure. Interpretation of DNA profiles is also dependent on well-developed theory of genetic inheritance, whereas interpretation in most branches of forensic science will have to be data driven (as is the case in machine learning in general including in biometric applications). The potential for transfer of new-paradigm knowledge and methods from DNA to other branches of forensic science is therefore limited.

Since around 2000, a growing number of researchers and practitioners in **forensic voice comparison** have developed and adopted methods for calculation of likelihood ratios based on acoustic measurements and statistical models, and methods for calibration

and validation of likelihood-ratio systems under casework conditions. This has included adoption of state-of-the-art machine-learning approaches to automatic speaker recognition (Lee et al., 2020; Matějka et al., 2020; Morrison, Enzinger, et al., 2020; Villalba et al., 2020; Weber et al., 2022). At present, however, only a minority of practitioners have adopted the new paradigm: In a survey of law-enforcement agencies in INTERPOL member countries (Morrison et al., 2016), among respondents who had forensic-voice-comparison capabilities, the reported rates of adoption of human-supervised-automatic approaches and numeric likelihood ratios were 33% and 23% respectively; however, even the latter have probably not adopted all elements of the new paradigm. In a survey including private practitioners (Gold & French, 2019), the reported rates of adoption of human-supervised-automatic approaches and numeric likelihood ratios were 41% and 13% respectively (up from 20% and 9% in Gold & French, 2011); however, inconsistent with the new paradigm, most reported combining human-supervised-automatic approaches with human-perception- and subjective-judgement-based approaches. Data in human-supervised-automatic approaches are continuously valued and have intrinsic within-source variability, a data structure shared with many other branches of forensic science. Compared to DNA, new-paradigm knowledge and methods from forensic voice comparison, including statistical models and calibration and validation procedures, should therefore be easier to transfer to and adapt for other branches of forensic science. Forensic voice comparison is, however, a relatively low-volume branch of forensic science, which limits the extent to which developments in forensic voice comparison are noticed and adopted by researchers and practitioners in other branches of forensic science.

Curran (2013) lamented that only 13% of laboratories surveyed used the likelihood-ratio framework for **glass evidence**, but this may be one of the highest rates of adoption after DNA. In many **other branches of forensic science**, including firearm examination, the rate of adoption of the likelihood-ratio framework by practitioners is near zero (Bali et al., 2020; Cole & Barno, 2020).

2.3.3 Only some elements of the new paradigm have been adopted as part of incremental change

Only some elements of the new paradigm have been adopted as part of incremental change. Although in the short term this may be a viewed as a step in the right direction, in the long term it may actually impede a paradigm shift.

> Just because it is a transition between incommensurables, the transition between competing paradigms cannot be made a step at a time, ... Like the gestalt switch, it must occur all at once (though not necessarily in an instant) or not at all. (Kuhn, 1962, p. 149)

Some practitioners **assign likelihood-ratio values based on subjective judgement**, and the values they assign are not subject to empirical calibration or empirical validation (see Risinger, 2013; Morrison & Thompson, 2017; Morrison, Ballantyne, Geoghegan, 2020). Some authors emphasize the logic of the likelihood-ratio framework and consider subjective assignment of likelihood ratio an acceptable end goal or consider it a step in the right direction, but such incremental steps potentially impede a paradigm shift which requires the abandonment of interpretation methods that are entirely dependent on subjective judgement.[6] In addition, placing an emphasis on subjectivist concepts of probability is detrimental to attempts to encourage practitioners to adopt methods based on relevant data, quantitative measurements, and statistical models, and to adopt empirical validation under casework conditions (Morrison, 2017).

The majority of **proposals to address cognitive bias** in forensic science (e.g., EWG, 2012; Stoel et al., 2015; Thompson et al., 2017; FSR, 2020a) **assume the continued use of human-perception- and subjective-judgement-based methods**. Although this may be necessary in the short term, it potentially impedes a paradigm shift to

[6] Our stance on where in the forensic-evaluation process use of subjective judgement is acceptable is more restrictive than that of some leaders in the field, e.g., Evett et al. (2017).

quantitative-measurement- and statistical-model-based methods.

Some practitioners **make use of systems based on quantitative measurements and statistical models, but do not empirically calibrate or validate** their system using data that reflect the relevant population and the conditions for the case, and/or **rather than directly report the output of the system, they use it as input to a subjective-judgement process** that also considers other information including from human-perception-based analyses (see Morrison & Thompson, 2017; Morrison, 2018a, 2018b). Such approaches are pernicious in that use of technology may give the false impression of scientific validity, and reaction against this may impede a paradigm shift that includes adoption of quantitative measurements and statistical models.

2.3.4 There is misunderstanding of the new paradigm and resistance to its adoption

As with all Kuhnian paradigm shifts, there is misunderstanding of the new paradigm and resistance to its adoption. **Some resistance is cultural**. People in general tend to prefer methods which involve greater human input even when validation results indicate that data-driven methods lead to better results, but over time they can come to accept the latter (Swofford & Champod, 2021). The cultures of some branches of forensic science seem to be especially resistant to the adoption of statistical-model-based methods and of validation (see Mnookin et al., 2011; Curran, 2013; Morrison, 2014; Morrison & Stoel, 2014; Morrison, Neumann, et al., 2020, 2021; Swofford et al., 2021).

There is a **belief that likelihood ratios are difficult to understand** (Bali et al. 2020; Swofford et al., 2021). Commonly occurring misunderstandings have even been given names, e.g., the “prosecutor’s fallacy” and the “defense attorney’s fallacy” (Thompson & Schurmann, 1987). There are many examples of legal rulings in which judges have misunderstood the meaning of a likelihood ratio (the England & Wales Court of Appeal 2010 ruling in *R v T* is an infamous example, see, e.g., Berger et al., 2011; Redmayne

et al., 2011; Morrison, 2012; Thompson, 2012). Results of empirical research on lay understanding of expressions of strength of evidence are mixed (Eldridge, 2019; Martire & Edmond, 2020).

Despite legal rulings and recommendations concerning the need for validation, **courts often do not understand empirical validation and its necessity**, and accept testimony based on forensic-evaluation methods that have not been validated under conditions reflecting those of the case under consideration, or even that have not been empirically validated at all (see Bernstein, 2013; Morrison, 2014, 2018a; Cooper, 2016; Edmond, 2018).

2.3.5 Research is often not informed by practice and has no impact on practice

Research is often not informed by practice and has no impact on practice. **Research that appears to be about forensic science may not actually be about solving real forensic-science problems**. For example, research may really be about a method in statistics or machine learning with an apparent forensic-science problem being used as an example of the application of that method.

> Research in forensic science is sorely needed, but it should address primarily forensic science questions—not questions relating to the application of chemistry, biology, statistics, or psychology. (Margot, 2011, p. 801)

Research that is not focused on solving real forensic science problems and is divorced from forensic practice may lead to academic papers but nothing else.

> it is critical that researchers and funding bodies understand the importance of conducting research that is informed by practice and can be translated into practical applications. (Roux & Weyermann, 2021, p. 2)

Even research that does address real forensic-science problems will fail to have impact unless it involves genuine collaboration in which researchers understand the demands

of practice, and in which practitioners are willing to embrace research-informed change (Curran, 2013).

2.3.6 It is difficult to obtain funding for evidential-forensic-science research

It is difficult to obtain funding for evidential-forensic-science research. **Few funding agencies have sustained funding targeted at forensic science**, and funding agencies seldom have panels of reviewers knowledgeable about evidential forensic science. Applications for funding for evidential-forensic-science research made to non-forensic-science-targeted calls are often rejected because reviewers do not understand the epistemology or value of forensic-science research. Applications are often rejected because their goals are to improve forensic science, which is an applied science, and **funding-agency criteria or reviewers often do not value applied science**.

> The larger scientific community must now come to the aid of our forensic colleagues in advocating both for: (i) the research and financial support that is so clearly needed to advance the field and (ii) the requirement for empirical testing that is so clearly needed to advance the cause of justice. ... Forensic scientists have long complained that their work is not always valued by their scientific colleagues because of its applied nature; it is time for the scientific community to move beyond that conceit. (Bell et al., 2018)

At the other extreme, when there are calls specific to forensic science, they usually **focus exclusively or primarily on short-term goals related to law-enforcement investigative applications rather than on courtroom-evidential applications** (investigative and evidential applications have very different requirements), and they usually focus on technology rather than on forensic inference.

> technology-oriented development ... often overrul[es] the importance of appropriate scientific reasoning to solve actual problems (Roux et al., 2021, p. 679)

Research calls requiring deliverables with a high technology readiness level (TRL) are not sources of funding for paradigm-shifting research.

In 2018, United Kingdom Research and Innovation (UKRI) informed the UK House of Lords Science and Technology Select Committee that UKRI had invested GBP 56M over 10 years in forensic-science research (less than 0.1% of UKRI's total budget), but on closer inspection most of the funding counted to obtain that figure was not for research projects that actually focused on (or even made any contribution to) forensic science: only about GBP 17M went to forensic-science focussed research, and GBP 15M of that went to TRL research, only GBP 2M to foundational research (HoL, 2019; Morgan & Levin, 2019). HoL (2019) recommended that UKRI "urgently and substantially increase the amount of dedicated funding allocated to forensic science" (§187), but (as of time or writing ~2.5 years after the publication of the HoL report) this has not (yet) happened.

2.3.7 There are genuine practical impediments to implementing the new paradigm

There are genuine practical impediments to implementing the new paradigm. Even if practitioners want to adopt the new paradigm, they will be unable to do so unless they are provided with the quantitative-measurement and statistical-modelling tools and the case-relevant data necessary to calculate likelihood ratios and validate systems under the conditions of the cases on which they work. Practitioners will also not be able to adopt the new paradigm unless they are provided with training on understanding the new paradigm in general and on how to implement it for the types of cases they work on.

2.4 Via progredi

Kuhn (1962):

> The transfer of allegiance from paradigm to paradigm is a conversion experience

> that cannot be forced. ... a generation is sometimes required to effect the change ... Conversions will occur a few at a time until, after the last holdouts have died, the whole profession will again be practicing under a single, but now a different, paradigm. (pp. 150–151)

Kuhnian paradigm shifts are not rapid and individuals cannot be forced to embrace the new paradigm, but our aim is to facilitate and thereby advance the adoption of the new paradigm. Our **strategy** is to work with researchers and practitioners who want to adopt the new paradigm, to work with them on addressing practical impediments to applying the new paradigm in casework, i.e.:

1. to provide researchers, practitioners, and lawyers with training leading to understanding of the new paradigm;

2. to collaborate with researchers and practitioners on building relevant databases and on developing and validating statistical models applicable in their particular branches of forensic science; and

3. to conduct research on how to present likelihood ratios and validation results so as to maximize understanding by laypeople, and thereby provide guidance to forensic practitioners on how to communicate forensic-evaluation results to legal-decision makers.

Part II of the present paper will focus on element 2 of the strategy. We will focus on elements 1 and 3 elsewhere. Element 2 of the strategy requires collaboration between researchers with expertise in forensic data science and researchers and practitioners with expertise in particular branches of forensic science. Academic publications are unlikely to convince practitioners to adopt the new paradigm, but other practitioners successfully applying the new paradigm are potentially convincing. In any branch of forensic science, the number of practitioners who initially want to adopt the new paradigm and who want to collaborate on this endeavour will almost certainly be a very

small minority, but it will be more productive to work with a small minority on developing practical solutions than to try to convince the majority of practitioners without providing practical solutions. Once the practical solutions are being used by the small minority, use of the new paradigm has the potential to spread. Even then, we do not expect adoption of the new paradigm to be rapid, but we do expect higher rates of adoption among newer practitioners and trainees, leading to a generational shift.

3 Calculating likelihood ratios and validating likelihood-ratio systems

3.1 Introduction

A new paradigm introduces new ways of thinking, which can introduce new problems to be solved, problems that may not even have been conceivable under the old paradigm. In the context of source attribution, one such problem is the need to take account not only of the similarity between the items of interest, but also of their typicality with respect to the relevant population. The relevant population is the population from which, in the context of the legal case, the item of questioned origin could conceivably have come had it not come from the known source.[7] The data used to train the likelihood-ratio models (particularly the model in the denominator) must be representative of the relevant population for the case, otherwise the calculated likelihood ratio will not answer the question of interest for the case.

Typicality with respect to the relevant population can be incorporated into the calculation of a likelihood ratio using either a specific-source or a common-source approach (Ommen & Saunders, 2021). Another approach, similarity-score-based

[7] The relevant population could be explicitly proposed by the defence, but in common-law systems the defence in a criminal trial is under no obligation to propose an alternative to the prosecution's proposition. A forensic practitioner usually has to decide what to adopt as the relevant population and communicate that decision to the court so that the court can later decide whether the forensic practitioner's decision was appropriate, i.e., will it result in a likelihood ratio that answers a question of interest for the court.

calculation of likelihood ratios does not properly incorporate typicality with respect to the relevant population. We discuss each of these approaches in §3.2 below.

Another problem introduced by the new paradigm is how to validate the performance of forensic-evaluation systems that output likelihood ratios. In a classification framework, a test-pair input is either "same source" or "different source" and the system being tested outputs either "same source" or "different source". The *miss rate* can be calculated as the proportion of "same source" inputs that elicited "different source" outputs, the *false-alarm rate* can be calculated as the proportion of "different source" inputs that elicited "same source" outputs, and the mean of the miss rate and false-alarm rate can be reported as the *classification-error rate*, a single-value summarizing the performance of the system. In a likelihood-ratio framework, however, the system being tested does not output a discrete classification, instead, it outputs a continuous likelihood-ratio value. What must be assessed in terms of performance is the gradient goodness of the output: Given a same-source input, a good output would be a likelihood-ratio value that is much larger than 1, a less good output would be a value that is only a little larger than 1, a bad output would be a value less than 1, and a worse output would be a value much less than 1. *Mutatis mutandis*, given a different-source input, a good output would be a value much less than 1. In §3.3, we describe a popular single-value metric for summarizing the performance of likelihood-ratio systems, the *log-likelihood-ratio cost* (C_{llr}, Brümmer and du Preez, 2006), and a popular graphical representation of likelihood-ratio validation results, a *Tippett plot* (Meuwly, 2001). We will use C_{llr} and Tippett plots when reporting the results of the empirical research described in Part II of the present paper.

3.2 Calculating likelihood ratios

3.2.1 Specific-source approach

A specific-source likelihood ratio answers the two-part question:

1. What is the probability of obtaining the measured properties of the item of questioned source if it came from the specific known source?

versus

2. What is the probability of obtaining the measured properties of the item of questioned source if it came not from the specific known source but from some other source selected at random from the relevant population?

A specific-source calculation has the form given in Equation (1), in which Λ is the likelihood ratio, $f(x|M)$ is a probability-density function, x_{q} is the feature (or feature vector) extracted from the questioned-source item, and M_{k} and M_{r} are the specific-known-source model and the relevant-population model respectively. The specific-known-source model is trained using features extracted from multiple items sampled from the specific known source, and the relevant-population model is trained using features extracted from multiple items sampled from the relevant population.

(1)

$$\Lambda = \frac{f\left(x_{\mathrm{q}}\middle|M_{\mathrm{k}}\right)}{f\left(x_{\mathrm{q}}\middle|M_{\mathrm{r}}\right)}$$

Equation (2) provides an example of a simple model for calculating specific-source likelihood ratios. The model assumes univariate Gaussian distributions in both the numerator and the denominator, and assumes that all sources have the same within-source variance σ_{w}^2. In Equation (2), $f(x|\mu,\sigma^2)$, is a Gaussian probability-density distribution (parametrized using mean and variance), μ_{k} and μ_{r} are the specific-known-source mean and the relevant-population mean respectively, and σ_{w}^2 and σ_{b}^2 are the within-source variance and between-source variance respectively.

(2)

$$\Lambda = \frac{f\left(x_{\mathrm{q}}\middle|\mu_{\mathrm{k}}, \sigma_{\mathrm{w}}^2\right)}{f\left(x_{\mathrm{q}}\middle|\mu_{\mathrm{r}}, \sigma_{\mathrm{w}}^2 + \sigma_{\mathrm{b}}^2\right)}$$

Figure 1 shows examples of the calculation of two specific-source likelihood ratios using Equation (2). The wider distribution represents the relevant-population model (denominator model), and the two peakier distributions represent two different specific-source models (numerator models). For these examples: $\mu_{\mathrm{r}} = 0$, $\sigma_{\mathrm{b}}^2 = 100$, and $\sigma_{\mathrm{w}}^2 = 1$; for the filled circles $\mu_{\mathrm{k}} = 0$ and $x_{\mathrm{q}} = -1$, and for the unfilled circles $\mu_{\mathrm{k}} = 20$ and $x_{\mathrm{q}} = 19$.[8] The resulting likelihood-ratio values, corresponding to the filled and unfilled circles respectively, are $0.24/0.040 = 6$ and $0.24/0.0066 = 36$.

[insert figure about here]

Figure 1. Examples of the calculation of specific-source likelihood ratios.

3.2.2 Common-source approach

A common-source likelihood ratio answers the two-part question:

1. What is the probability of obtaining the measured properties of the items of questioned- and known-source if they both came from the same source (a source selected at random from the relevant population)?

   versus

2. What is the probability of obtaining the measured properties of the items of questioned- and known-source if they each came from a different source (each a source selected at random from the relevant population)?

[8] The values in these examples, and in other examples below, were chosen purely for illustrative purposes.

A common-source calculation has the form given in Equation (3), in which Λ is the likelihood ratio, $f(x_{\mathrm{q}}, x_{\mathrm{k}}|M)$ is a joint probability-density function, x_{q} and x_{k} are the features (or feature vectors) extracted from the questioned- and known-source items respectively, and M_{s} and M_{d} are the same-source and different-source models respectively.

(3)

$$\Lambda = \frac{f(x_{\mathrm{q}}, x_{\mathrm{k}}|M_{\mathrm{s}})}{f(x_{\mathrm{q}}|M_{\mathrm{d}})f(x_{\mathrm{k}}|M_{\mathrm{d}})}$$

Equation (4) provides an example of a simple model for calculating common-source likelihood ratios.[9] The model assumes univariate Gaussian distributions in both the numerator and the denominator, and assumes that all sources have the same within-source variance σ_{w}^2. The numerator of Equation (4) integrates over all possible values for source means given the between-source distribution, with the constraint that x_{q} and x_{k} come from the same source. The denominator of Equation (4) integrates over all possible values for source means given the between-source distribution, but does so independently for x_{q} and for x_{k}. The solutions to the integrals can be expressed as bivariate Gaussian distributions in which for the same-source model (the numerator model) the covariances equal the between-source variance, σ_{b}^2, but for the different-source model (the denominator model) the covariances are zero, 0. This reflects the logic that if a source is selected at random from the population and the mean of the source is high, then two items selected at random from that source will both be expected to have high values. Likewise, if the mean of the source is low, the values of both items will be expected to be low. In general, the values of two items selected from the same

[9] This is the univariate version of the model described Aitken & Lucy (2014) as the "multivariate normal (MVN) procedure", and in the automatic-speaker-recognition and forensic-voice-comparison literature (e.g., Prince & Elder, 2007; Kenny, 2010; Brümmer & de Villiers, 2010; Sizov et al., 2014; Morrison, Enzinger, et al., 2020) as the "two-covariance model" for "probabilistic linear discriminant analysis (PLDA)".

source are expected to be correlated. In contrast, if two sources are selected at random from the population and the mean of one is high, there is no expectation that the mean of the other source will also be high. The mean of the other source is more likely to the average, and it is equally likely to be low as to be high. The values of two items each selected from a different source are not expected to be correlated.

(4)

$$\Lambda = \frac{\int f\left(x_{\mathrm{q}}\middle|\mu_i,\sigma_{\mathrm{w}}^2\right) f\left(x_{\mathrm{k}}\middle|\mu_i,\sigma_{\mathrm{w}}^2\right) f\left(\mu_i\middle|\mu_{\mathrm{r}},\sigma_{\mathrm{b}}^2\right) d\mu_i}{\int f\left(x_{\mathrm{q}}\middle|\mu_i,\sigma_{\mathrm{w}}^2\right) f\left(\mu_i\middle|\mu_{\mathrm{r}},\sigma_{\mathrm{b}}^2\right) d\mu_i \int f\left(x_{\mathrm{k}}\middle|\mu_j,\sigma_{\mathrm{w}}^2\right) f\left(\mu_j\middle|\mu_{\mathrm{r}},\sigma_{\mathrm{b}}^2\right) d\mu_j}$$

$$= \frac{f\left(\begin{bmatrix} x_{\mathrm{q}} \\ x_{\mathrm{k}} \end{bmatrix}\middle|\begin{bmatrix} \mu_{\mathrm{r}} \\ \mu_{\mathrm{r}} \end{bmatrix}, \begin{bmatrix} \sigma_{\mathrm{w}}^2+\sigma_{\mathrm{b}}^2 & \sigma_{\mathrm{b}}^2 \\ \sigma_{\mathrm{b}}^2 & \sigma_{\mathrm{w}}^2+\sigma_{\mathrm{b}}^2 \end{bmatrix}\right)}{f\left(x_{\mathrm{q}}\middle|\mu_{\mathrm{r}},\sigma_{\mathrm{w}}^2+\sigma_{\mathrm{b}}^2\right) f\left(x_{\mathrm{k}}\middle|\mu_{\mathrm{r}},\sigma_{\mathrm{w}}^2+\sigma_{\mathrm{b}}^2\right)}$$

$$= \frac{f\left(\begin{bmatrix} x_{\mathrm{q}} \\ x_{\mathrm{k}} \end{bmatrix}\middle|\begin{bmatrix} \mu_{\mathrm{r}} \\ \mu_{\mathrm{r}} \end{bmatrix}, \begin{bmatrix} \sigma_{\mathrm{w}}^2+\sigma_{\mathrm{b}}^2 & \sigma_{\mathrm{b}}^2 \\ \sigma_{\mathrm{b}}^2 & \sigma_{\mathrm{w}}^2+\sigma_{\mathrm{b}}^2 \end{bmatrix}\right)}{f\left(\begin{bmatrix} x_{\mathrm{q}} \\ x_{\mathrm{k}} \end{bmatrix}\middle|\begin{bmatrix} \mu_{\mathrm{r}} \\ \mu_{\mathrm{r}} \end{bmatrix}, \begin{bmatrix} \sigma_{\mathrm{w}}^2+\sigma_{\mathrm{b}}^2 & 0 \\ 0 & \sigma_{\mathrm{w}}^2+\sigma_{\mathrm{b}}^2 \end{bmatrix}\right)}$$

Figure 2 shows examples of the calculation of two common-source likelihood ratios using Equation (4). In Figure 2a the univariate distribution represents the relevant-population model with $\mu_{\mathrm{r}} = 0$, and $\sigma_{\mathrm{r}}^2 = \sigma_{\mathrm{w}}^2 + \sigma_{\mathrm{b}}^2$ ($\sigma_{\mathrm{b}}^2 = 100$, and $\sigma_{\mathrm{w}}^2 = 1$), and the pairs of circles represent pairs of x_{q} and x_{k} feature values. For the filled circles $x_{\mathrm{q}} = -1$ and $x_{\mathrm{k}} = 1$, and for the unfilled circles $x_{\mathrm{q}} = 19$ and $x_{\mathrm{k}} = 21$. Figure 2b shows the projection of the x_{q} and x_{k} feature values into the two-dimensional space of Equation (4), and shows the peakier same-source model (the numerator model) and the flatter different-source model (the denominator model). The resulting likelihood-ratio values, corresponding to the filled and unfilled circles respectively, are $(41 \times 10^{-4})/(16 \times 10^{-4}) = 2.6$ and $(56 \times 10^{-5})/(3.0 \times 10^{-5}) = 19$.

[insert figure about here]

Figure 2. Examples of the calculation of common-source likelihood ratios.

Although the example specific-source and common-source models above assume univariate Gaussian distributions, both specific-source and common-source models can fit more complex multivariate distributions. The latter can potentially exploit more useful information resulting in better performance, but they require larger amounts of data in order to estimate larger numbers of parameter values. In forensic casework, the amount of case-relevant data is often limited. Solutions may include use of dimension-reduction techniques, extraction of features using functional data analysis, and use of embeddings from deep neural networks (DNNs) as feature vectors.

3.2.3 Similarity-score approach

Across multiple branches of forensic science, there have been repeated proposals to calculate likelihood ratios using similarity-score-based approaches (see Morrison & Enzinger, 2018). Such approaches may be motivated by situations in which a specific-source approach cannot be adopted because the data only allow for a comparison between one feature vector and one other feature vector (and this constraint may be due to casework conditions, hence collecting more data may not be a possible solution). They may also be motivated by the difficulties of trying to fit potentially complex distributions in high-dimensional spaces using limited amounts of training data. Instead of fitting models to feature vectors, as for the specific-source and common-source models described above, models are fitted to scores which quantify degree of similarity (or, inversely, degree of difference) between pairs of items. Similarity-scores are scalar values that can be based, for example, on Manhattan distance, Euclidian distance, or a

correlation coefficient.[10] Similarity-scores are calculated for pairs of items that are known to come from the same source and for pairs of items known to come from different sources, resulting in a set of same-source scores and a set of different-source scores. The latter scores are used to train a model of the form given in Equation (5), in which $\delta(x_{\mathrm{q}}, x_{\mathrm{k}})$ is the function that calculates the score, and $M_{\delta,\mathrm{s}}$ and $M_{\delta,\mathrm{d}}$ are univariate models trained on same-source scores and different-source scores respectively.

(5)

$$\Lambda = \frac{f(\delta(x_{\mathrm{q}}, x_{\mathrm{k}})|M_{\delta,\mathrm{s}})}{f(\delta(x_{\mathrm{q}}, x_{\mathrm{k}})|M_{\delta,\mathrm{d}})}$$

A similarity-score-based likelihood ratio answers the two-part question:

1. What is the probability of obtaining the measured degree of similarity between the items of questioned- and known-source if they both came from the same source (a source selected at random from the relevant population)?

   versus

2. What is the probability of obtaining the measured degree of similarity between the items of questioned- and known-source if they each came from a different source (each a source selected at random from the relevant population)?

It may not be immediately obvious, but similarity-score-based approaches do not properly account for typicality with respect to the relevant population and are therefore

[10] Some published literature has failed to distinguish between the calculation of likelihood ratios based on similarity-scores and the calibration of uncalibrated likelihood ratios (see Morrison, 2013, for an introduction to the latter). The confusion may stem from the fact that uncalibrated log likelihood ratios have usually been called "scores" (and the calibration process has been called "score to likelihood-ratio conversion"), but these scores are not similarity-scores, they are scores that take account of both similarity and typicality with respect to the relevant population.

not appropriate for calculating likelihood ratios addressing questions of interest for a court. We briefly demonstrate that similarity-scores do not properly account for typicality with respect to the relevant population (more extensive arguments and demonstrations are presented in Morrison & Enzinger, 2018; Neumann & Ausdemore, 2020; and Neumann et al., 2020). In Figure 2a the distance between the two filled circles and the distance between the two non-filled circles is the same, 2 in each case, hence the two pairs of circles will both have the same similarity-score value. The two curves in Figure 3 represent Weibull distributions fitted to same-source scores and to different-source scores, which were obtained by drawing Monte Carlo samples of pairs of same-source feature values and pairs of different-source feature values from the population distribution (consisting of within- and between-source Gaussian distributions with $\mu_\mathrm{r} = 0$, $\sigma_\mathrm{b}^2 = 100$, and $\sigma_\mathrm{w}^2 = 1$), and using unsigned difference as the score function, $\delta(x_\mathrm{q}, x_\mathrm{k}) = |x_\mathrm{q} - x_\mathrm{k}|$. Since the two pairs of circles in Figure 2a have the same score value $\delta(x_\mathrm{q}, x_\mathrm{k}) = 2$, when score-based calculation of likelihood ratios is applied (as graphically illustrated in Figure 3) they both result in the same likelihood-ratio value, which in this example is $0.18/0.058 = 3.1$. Note, however, that in the feature space in Figure 2a, the probability of obtaining the pair of filled-circle feature values if each came from one of two different sources each selected at random from the population is higher than the probability of obtaining the pair of non-filled-circle feature values if each of them came from one of two different sources each selected at random from the population, because the former pair are more typical than the latter pair. Selecting two sources at random, the probability that they will both be in the middle of the distribution is relatively high, whereas the probability that they will both be on the same tail of the distribution is very low. Hence, the likelihood-ratio value associated with the unfilled circles should be higher than the likelihood-ratio value associated with the filled circles, which was the case when the common-source approach was used (19 versus 2.6), but was not the case when the similarity-score-based approach was used (both 3.1), *quod erat demonstandum*: similarity-score-based approaches do not properly account for typicality with respect to the relevant

population.

[insert figure about here]

Figure 3. Example of the calculation of a similarity-score-based likelihood ratio.

Some published research reports have cited publications such as Morrison & Enzinger (2018), Neumann & Ausdemore (2020), and Neumann et al. (2020), but have then proceeded to use similarity-score-based approaches anyway. Sometimes, they characterize the use of similarity-score-based approaches as entailing "some" loss of information, but we consider the loss of information about typicality with respect to the relevant population to be a loss of essential information, hence we take the position that the use of similarity-score-based approaches are not appropriate for evaluating strength of forensic evidence.

3.3 Validating the performance of a likelihood-ratio system

3.3.1 Protocol

Protocols for validating systems that output likelihood ratios, and metrics and graphics for representing the results of such validations are described in: Morrison (2011); Meuwly et al. (2017); Ramos et al. (2020); Morrison, Enzinger, et al. (2021).

In order to empirically validate the performance of a forensic-evaluation system that outputs likelihood ratios, the tester must input to the system pairs of test items for which it is known that the two items come from the same source and pairs of test items for which it is known that the two items come from different sources. Each pair of items must be sampled from the relevant population, and must reflect the conditions of the questioned-source and known-source items in the case. If there is a mismatch in the conditions of the questioned-source and known-source items, one member of each test pair must reflect the conditions of the questioned-source item and the other must reflect

the conditions of the known-source item. If the test pairs do not represent the relevant population and/or do not reflect the conditions of the case, then the validation results will not be indicative of the expected performance of the system in the context of the case. The tester inputs a set of same-source pairs and obtains a set of same-source likelihood-ratio values as outputs, and inputs a set of different-source pairs and obtains a set of different-source likelihood-ratio values as outputs.

The system being tested must not be told the truth as to which input pairs are same source and which input pairs are different source. If the system automatically extracts features and passes the feature values to statistical models, the validation process can be automated, and the forensic practitioner can act as the tester.

3.3.2 Log-likelihood-ratio cost (C_{llr})

As previously stated in §3.1, given a same-source input, a good output would be a likelihood-ratio value that is much larger than 1, a less good output would be a value that is only a little larger than 1, a bad output would be a value less than 1, and a worse output would be a value much less than 1. *Mutatis mutandis*, given a different-source input, a good output would be a value much less than 1. A metric that captures this gradient goodness is C_{llr}, which is calculated as in Equation (6), in which Λ_s and Λ_d are likelihood-ratio outputs corresponding to same-source and different-source input pairs respectively, and N_s and N_d are the number of same-source and different-source input pairs respectively. C_{llr} is equivalent to the deviance statistic with equal priors.

(6)

$$C_{\mathrm{llr}} = \frac{1}{2}\left(\frac{1}{N_{\mathrm{s}}}\sum_{i=1}^{N_{\mathrm{s}}}\log_2\left(1+\frac{1}{\Lambda_{\mathrm{s}_i}}\right)+\frac{1}{N_{\mathrm{d}}}\sum_{j=1}^{N_{\mathrm{d}}}\log_2\left(1+\Lambda_{\mathrm{d}_j}\right)\right)$$

Figure 4 plots the logarithmic penalty functions for same-source likelihood-ratio outputs (the function within the leftmost summation in Equation (6)) and for different-

source likelihood-ratio outputs (the function within the rightmost summation in Equation (6)). The *x*-axis of Figure 4 is scaled in $\log_{10}$ likelihood ratios. If the value of a same-source log likelihood ratio is much greater than 0 it receives a small penalty value, but if its value is lower it receives a higher penalty value. If the value of a different-source log likelihood ratio is much less than 0 it receives a small penalty value, but if its value is higher it receives a higher penalty value. C_{llr} is calculated as the mean of the penalty values with equal weight given to the set of same-source penalty values and the set of different-source penalty values.

[insert figure about here]

Figure 4. Penalty functions used in the calculation of C_{llr}.

As with classification-error-rate values, lower C_{llr} values indicate better performance, and C_{llr} values cannot be less than 0. A system that gave no useful information and always responded with a likelihood ratio of 1, irrespective of the input, would have a C_{llr} value of 1. A system with a C_{llr} of less than 1 is providing useful information. C_{llr} values substantially greater than 1 can be produced by uncalibrated or miscalibrated systems, but this can be resolved by appropriately calibrating the system.

C_{llr} was introduced by Brümmer & du Preez (2006) in the context of automatic speaker recognition. One of the first forensic-science papers to use C_{llr} was González-Rodríguez et al. (2007). Its use is recommended in the *Consensus on validation of forensic voice comparison* (Morrison, Enzinger, et al., 2021).

3.3.3 Tippett plot

Tippett plots consist of plots of the empirical cumulative probability distributions of the same-source likelihood-ratio values and of the different-source likelihood-ratio values. The tradition is to plot lines joining the data points rather than to plot the data

points themselves. Examples are shown in Figure 5. The *y*-axis values corresponding to the curves rising to the right give the proportion of same-source test results with log likelihood-ratio values less than or equal to the corresponding value on the *x*-axis. The *y*-axis values corresponding to the curves rising to the left give the proportion of different-source test results with log likelihood-ratio values greater than or equal to the corresponding value on the *x*-axis. In general, shallower curves with greater separation between the two curves indicates better performance. Tippett plots give an indication of the range of possible likelihood-ratio values that the system could generate under the test conditions, and can also reveal problems such as bias in the output. In Figure 5(a) the separation between the same-source and different-source curves is small and the system is clearly biased – both the same-source and different-source curves are too far to the right. The C_{llr} value corresponding to these results is 1.07. Figure 5(b) shows the results of a validation of a better performing system. This Tippett plot has somewhat greater separation between the same-source and different-source curves and the results are not obviously biased. The C_{llr} value corresponding to these results is 0.70. The system that generated the results in Figure 5(b) was actually the same as the system that generated the results in Figure 5(a) except that the Figure 5(b) system included a calibration stage whereas the Figure 5(a) system did not. Figure 5(c) shows the results of a validation of a system with substantially better performance – the same-source and different-source curves have greater separation and are shallower. The C_{llr} value corresponding to these results is 0.31.

[insert figure about here]

Figure 5. Example Tippett plots.

Tippett plots (i.e., plotting both same-source and different-source empirical cumulative distributions on the same plot) were introduced by Meuwly (2001) in the context of forensic voice comparison. Their use is recommended in the *Consensus on validation*

of forensic voice comparison (Morrison, Enzinger, et al., 2021).

4 Preview of Part II

Element 2 of our strategy (§2.4) is to collaborate with researchers and practitioners on building relevant databases and on developing and validating statistical models applicable in their particular branches of forensic science.

Forensic firearm examination is a branch of forensic science in which the new paradigm has so far made almost no progress. A common task in forensic firearm examination is forensic comparison of fired cartridge cases. In Part II of the present paper we provide an example of the application of element 2 of our strategy. In Part II:

1. We provide an overview of current practice in forensic comparison of fired cartridge cases.

2. We provide a brief review of published research on feature-extraction and statistical-modelling methods applied to forensic comparison of fired cartridge cases.

3. We describe the building of a database and the development and validation of feature-extraction and statistical-modelling methods for calculating likelihood ratios for forensic comparison of fired cartridge cases.

Part II focuses particularly on the problem of feature extraction, i.e., what are the best features to extract and what region of the base of a fired cartridge case is it best to extract the features from?

Part II is intended to provide an example of the application of the new paradigm that can potentially be copied in other branches of forensic science. The authors of Part II include researchers with expertise in machine learning, forensic data science, and

forensic firearm examination. The research is conducted in collaboration with forensic practitioners who have contributed advice and provided the fired cartridge cases used to build the database.

Part II describes a first attempt to develop a feature-based statistical model for calculating likelihood ratios from 3D digital images of the bases of fired cartridge cases. We will use the common-source approach for calculating likelihood ratios (§3.2.2). An advantage of the common-source approach is that it can be applied when there is a single questioned-source feature vector and a single known-source feature vector, whereas the specific-source approach requires multiple known-source feature vectors (e.g., each extracted from a different item sampled from the same specific known source) in order to train the specific-source model. This means that the common-source approach can also be applied when, rather than having a questioned-source item and one or more known-source items, there are only questioned-source items to be compared with one another: multiple questioned-source items cannot be assumed to have come from the same source, so a specific-source model cannot be trained using multiple feature vectors each obtained from a different questioned-source item. The statistical modelling process will be based on the back-end of a pipeline commonly used in human-supervised-automatic approaches to forensic voice comparison. The features will be extracted using functional data analysis, i.e., they will be values of functions fitted to the 3D-image data. In this research, we will focus particularly on feature extraction, and will test several different feature sets. The validation will make use of protocols, metrics, and graphics commonly used in human-supervised-automatic approaches to forensic voice comparison (§3.3).

The database is described in Part II. The current size of the database is relatively small (the COVID-19 pandemic slowed data collection), hence we have kept the statistical-modelling process relatively parsimonious. We plan to continue increasing the size of the database and then conduct additional research to develop and validate more complex models that could potentially have better performance. After the latter stage,

we will collaborate with practitioners on field testing a prototype forensic-evaluation system.

5 Conclusion

A paradigm shift in evaluation of forensic evidence is ongoing. The shift is away from methods based on human perception and subjective judgement, to methods based on relevant data, quantitative measurements, and statistical models; methods that are transparent and reproducible, that are intrinsically resistant to cognitive bias, that use the logically correct framework for interpretation of evidence (the likelihood-ratio framework), and that are empirically validated under casework conditions. The new paradigm can be called *forensic data science*.

Some branches of forensic science, such as forensic voice comparison, are more advanced in the paradigm shift than others. Knowledge gained in advancing the paradigm shift in forensic voice comparison can assist in advancing the paradigm shift in other branches of forensic science. Statistical models used in forensic voice comparison can even be transferred and adapted for use in other branches of forensic science.

Our strategy for advancing the paradigm shift requires collaboration between researchers with expertise in forensic data science and researchers and practitioners with expertise in particular branches of forensic science. Our strategy is to work with researchers and practitioners who want to adopt the new paradigm. Our strategy is to work with them on addressing practical impediments to applying the new paradigm in casework, i.e.:

1. to provide researchers, practitioners, and lawyers with training leading to understanding of the new paradigm;

2. to collaborate with researchers and practitioners on building relevant databases and on developing and validating statistical models applicable in their particular branches of forensic science; and

3. to conduct research on how to present likelihood ratios and validation results so as to maximize understanding by laypeople, and thereby provide guidance to forensic practitioners on how to communicate forensic-evaluation results to legal-decision makers.

In any branch of forensic science, the number of practitioners who initially want to adopt the new paradigm and who want to collaborate on this endeavour will almost certainly be very small, but it will be more productive to work with a small minority on developing practical solutions than to try to convince the majority of practitioners without providing practical solutions. Once the practical solutions are being used by the small minority, use of the new paradigm has the potential to spread. Even then, we do not expect adoption of the new paradigm to be rapid, but we do expect higher rates of adoption among newer practitioners and trainees, leading to a generational shift.

In Part II of the present paper, we focus on building a relevant database and developing and validating statistical models for forensic comparison of fired cartridge cases, a common tasks in forensic firearm examination, a branch of forensic science in which the new paradigm has so far made almost no progress. Part II focuses particularly on the problem of feature extraction, i.e., what are the best features to extract and what region of the base of a fired cartridge case is it best to extract the features from? This is a first step in attempting to advance the paradigm shift in that branch of forensic science.

6 References

Aitken, C.G.G., Berger, C.E.H., Buckleton, J.S., Champod, C., Curran, J.M., Dawid, A.P., Evett, I.W., Gill, P., González-Rodríguez, J., Jackson, G., Kloosterman, A., Lovelock, T., Lucy, D., Margot, P., McKenna, L., Meuwly, D., Neumann, C., Nic Daéid, N., Nordgaard, A., Puch-Solis, R., Rasmusson, B., Redmayne, M., Roberts, P., Robertson, B., Roux, C., Sjerps, M.J., Taroni, F., Tjin-A-Tsoi, T., Vignaux, G.A., Willis, S.M., Zadora, G. (2011) Expressing evaluative opinions: A position statement. *Science & Justice*, 51, 1–2. http://dx.doi.org/10.1016/j.scijus.2011.01.002

Aitken, C.G.G., Lucy, D. (2004) Evaluation of trace evidence in the form of multivariate data. *Applied Statistics*, 53, 109–122. http://dox.doi.org/10.1046/j.0035-9254.2003.05271.x [Corrigendum: (2004) 53, 665–666. http://dox.doi.org/10.1111/j.1467-9876.2004.02031.x]

Aitken, C.G.G., Roberts, P., Jackson, G. (2010) *Fundamentals of Probability and Statistical Evidence in Criminal Proceedings: Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses*. London, UK: Royal Statistical Society. (Available from https://rss.org.uk/news-publication/publications/law-guides/)

Association of Forensic Science Providers (2009) Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice*, 49, 161–164. http://dx.doi.org/10.1016/j.scijus.2009.07.004

Bali, A.S., Edmond, G., Ballantyne, K.N., Kemp, R.I., Martire, K.A. (2020) Communicating forensic science opinion: An examination of expert reporting practices. *Science & Justice*, 60, 216–224. https://doi.org/10.1016/j.scijus.2019.12.005

889 Ballantyne, K., Bunford, J., Found, B., Neville, D., Taylor, D., Wevers, G., Catoggio,
890 D. (2017) *An Introductory Guide to Evaluative Reporting*. National Institute of
891 Forensic Science of the Australia New Zealand Policing Advisory Agency.
892 (Available from http://www.anzpaa.org.au/forensic-science/our-
893 work/projects/evaluative-reporting)

894 Basu, N., Bolton-King R., Morrison, G.S. (2022) Advancing a paradigm shift in
895 evaluation of forensic evidence – Part II: Feature-extraction methods for forensic
896 comparison of fired cartridge cases. (Preprint available from
897 http://firearms.forensic-data-science.net/)

898 Bell, S., Sah, S., Albright, T.D., Gates, S.J., Denton M.B., Casadevall, A. (2018) A
899 call for more science in forensic science. *Proceedings of the National Academy*
900 *of Sciences*, 115, 4541–4544; https://doi.org/10.1073/pnas.1712161115

901 Berger, C.E.H., Buckleton, J., Champod, C., Evett, I.W., Jackson, G. (2011) Evidence
902 evaluation: A response to the Court of Appeal judgment in R v T. *Science &*
903 *Justice*, 51, 43–49. http://dx.doi.org/10.1016/j.scijus.2011.03.005

904 Bernstein, D.E. (2013) The misbegotten judicial resistance to the Daubert revolution.
905 *Notre Dame Law Review*, 89, 27–70. (Available from
906 http://ndlawreview.org/publications/archives/volume-89/issue-1/)

907 Brümmer, N., du Preez, J. (2006) Application independent evaluation of speaker
908 detection. *Computer Speech and Language*, 20, 230–275.
909 https://doi.org/10.1016/j.csl.2005.08.001

910 Brümmer, N., de Villiers, E. (2010) The speaker partitioning problem. In Proceedings
911 of Odyssey 2010: The speaker and language recognition workshop, pp. 194–
912 201. (Available from https://www.isca-
913 speech.org/archive_open/odyssey_2010/od10_034.html)

914 Cole, S.A. (2006) Is fingerprint identification valid? Rhetorics of reliability in
915 fingerprint proponents' discourse. *Law & Policy*, 28, 109–135.
916 https://doi.org/10.1111/j.1467-9930.2005.00219.x

917 Cole, S.A. (2009) Forensics without uniqueness, conclusions without
918 individualization: The new epistemology of forensic identification. *Law,*
919 *Probability and Risk*, 8, 233−255. https://doi.org/10.1093/lpr/mgp016

920 Cole, S.A. (2014) Individualization is dead, long live individualization! Reforms of
921 reporting practices for fingerprint analysis in the United States. *Law, Probability*
922 *and Risk*, 13, 117−150. https://doi.org/10.1093/lpr/mgt014

923 Cole, S.A., Barno, M. (2020) Probabilistic reporting in criminal cases in the United
924 States: A baseline study. *Science & Justice*, 60, 406–414.
925 https://doi.org/10.1016/j.scijus.2020.06.001

926 Cooper, G.S., Meterko, V. (2019) Cognitive bias research in forensic science: A
927 systematic review. *Forensic Science International*, 297,35–46.
928 https://doi.org/10.1016/j.forsciint.2019.01.016

929 Cooper, S.L. (2016) Forensic science identification evidence: Tensions between law
930 and science. *Journal of Philosophy, Science & Law*, 16(2), 1–35.
931 https://doi.org/10.5840/jpsl20161622

932 Curran, J.M. (2013) Is forensic science the last bastion of resistance against statistics?
933 *Science & Justice*, 53, 251–252. http://dx.doi.org/10.1016/j.scijus.2013.07.001

934 Edmond, G. (2018) Re-assessing reliability. In *Forensic Science Evidence and Expert*
935 *Witness Testimony* (eds P. Roberts, M. Stockdale), pp. 71–105. Cheltenham,
936 UK: Elgar.

Edmond, G., Towler, A., Growns, B., Ribeiro, G., Found, B., White, D., Ballantyne, K., Searston, R.A., Thompson, M.B., Tangen, J.M., Kemp, R.I., Martire, K.A. (2017) Thinking forensics: Cognitive science for forensic practitioners. *Science & Justice*, 57, 144–154. http://dx.doi.org/10.1016/j.scijus.2016.11.005

Evett, I.W., Berger, C.E.H., Buckleton, J.S., Champod, C., Jackson, G. (2017) Finding the way forward for forensic science in the US – A commentary on the PCAST report. *Forensic Science International*, 278, 16–23. http://dx.doi.org/10.1016/j.forsciint.2017.06.018

Expert Working Group on Human Factors in Latent Print Analysis (2012) *Latent Print Examination and Human Factors: Improving the Practice through a Systems Approach*. Gaithersburg, MD: National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.7842

Foreman, L.A., Champod, C., Evett, I.W., Lambert, J.A., Pope, S. (2003) Interpreting DNA evidence: A review. *International Statistical Review*, 71, 473–495. http://dx.doi.org/10.1111/j.1751-5823.2003.tb00207.x

Forensic Science Regulator (2020a) Guidance: Cognitive Bias Effects Relevant to Forensic Science Examinations (FSR-G-217 Issue 2). (Available from https://www.gov.uk/government/publications/cognitive-bias-effects-relevant-to-forensic-science-examinations)

Forensic Science Regulator (2020b). Guidance: Validation (FSR-G-201 Issue 2). (Available from https://www.gov.uk/government/publications/forensic-science-providers-validation)

Forensic Science Regulator (2021) Codes of Practice and Conduct: Development of Evaluative Opinions (FSR-C-118 Issue 1). (Available from https://www.gov.uk/government/publications/development-of-evaluative-opinions)

963 Found, B. (2015) Deciphering the human condition: The rise of cognitive forensics.
964 *Australian Journal of Forensic Sciences*, 47, 386–401.
965 http://dx.doi.org/10.1080/00450618.2014.965204

966 Gold, E., French, J.P. (2011) International practices in forensic speaker comparison.
967 *International Journal of Speech, Language and the Law*, 18, 143–152.
968 http://dx.doi.org/10.1558/ijsll.v18i2.293

969 Gold, E., French, J.P. (2019) International practices in forensic speaker comparison:
970 Second survey. *International Journal of Speech, Language and the Law*, 26, 1–
971 20. https://doi.org/10.1558/ijsll.38028

972 Eldridge, H. (2019). Juror comprehension of forensic expert testimony: A literature
973 review and gap analysis. *Forensic Science International: Synergy*, 1, 24–34.
974 https://doi.org/10.1016/j.fsisyn.2019.03.001

975 House of Lords Science and Technology Select Committee (2019) Forensic Science
976 and the Criminal Justice System: A Blueprint for Change. (Available from
977 https://publications.parliament.uk/pa/ld201719/ldselect/ldsctech/333/333.pdf)

978 Jackson, G. (2009) Understanding forensic science opinions. In *Handbook of*
979 *Forensic Science* (eds J. Fraser, R. Williams), pp. 419–445. Cullompton, UK:
980 Willan. https://doi.org/10.4324/9781843927327

981 Kafadar, K., Stern, H., Cuellar, M., Curran, J., Lancaster, M., Neumann, C.,
982 Saunders, C., Weir, B., Zabell, S. (2019) American Statistical Association
983 Position on Statistical Statements for Forensic Evidence. (Available from
984 https://www.amstat.org/asa/files/pdfs/POL-ForensicScience.pdf)

985 Kenny, P. (2010) Bayesian speaker verification with heavy tailed priors. In
986 Proceedings of Odyssey 2010: The Speaker and Language Recognition
987 Workshop, paper 014. (Available from https://www.isca-
988 speech.org/archive_open/odyssey_2010/od10_014.html)

989 Kaye, D.H. (2015) Presenting forensic identification findings: The current situation.
990 In Communicating the Results of Forensic Science Examinations: Final
991 Technical Report for NIST Award Number 70NANB12H014 (eds C. Neumann,
992 A. Ranadive, D.H. Kaye), pp. 12–30. (Available from
993 https://ssrn.com/abstract=2690899)

994 Koehler, J.J. (2017) Forensics or fauxrensics? Ascertaining accuracy in the forensic
995 sciences. *Arizona State Law Journal*, 49(4), 1369–1416. (Available from
996 https://arizonastatelawjournal.org/2018/02/07/forensics-or-fauxrensics-
997 ascertaining-accuracy-in-the-forensic-sciences/)

998 Kuhn, T.S. (1962) *The Structure of Scientific Revolutions*. Chicago IL: University of
999 Chicago Press.

1000 Lee, K.A., Yamamoto, H., Okabe, K., Wang, Q., Guo, L., Koshinaka, T., Zhang, J.,
1001 Shinoda, K. (2020) NEC-TT System for mixed-bandwidth and multi-domain
1002 speaker recognition. *Computer Speech, Language*, 61, article 101033.
1003 https://doi.org/10.1016/j.csl.2019.101033

1004 Margot, P. (2011) Commentary on the need for a research culture in the forensic
1005 sciences. *UCLA Law Review*, 58, 795–801. (Available from
1006 https://www.uclalawreview.org/commentary-on-the-need-for-a-research-culture-
1007 in-the-forensic-sciences-3-2/)

Matějka, P., Plchot, O., Glembek, O., Burget, L., Rohdin, J., Zeinali, H., Mošner, L., Silnova, A., Novotný, O., Diez, M., Černocký, J.H. (2020) 13 years of speaker recognition research at BUT, with longitudinal analysis of NIST SRE. *Computer Speech & Language*, 63, article 101035. https://doi.org/10.1016/j.csl.2019.101035

Meuwly, D. (2001) Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique. Doctoral dissertation, University of Lausanne. (Available from https://www.unil.ch/files/live/sites/esc/files/shared/These.Meuwly.pdf)

Meuwly, D., Ramos, D., Haraksim, R. (2017) A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic Science International*, 276, 142–153. http://dx.doi.org/10.1016/j.forsciint.2016.03.048

Mnookin, J.L., Cole, S.A., Dror, I.E., Fisher, B.A.J., Houck, M.M., Inman, K., Kaye, D.H., Koehler, J.J., Langenburg, G., Risinger, D.M., Rudin, N., Siegel, J., Stoney, D.A. (2011) The need for a research culture in the forensic sciences. *UCLA Law Review*, 58, 725–777. (Available from https://www.uclalawreview.org/the-need-for-a-research-culture-in-the-forensic-sciences-2/)

Morgan, R.M., Levin, E.A. (2019) A crisis for the future of forensic science: Lessons from the UK of the importance of epistemology for funding research and development. *Forensic Science International: Synergy*, 1, 243–252. https://doi.org/10.1016/j.fsisyn.2019.09.002

Morrison, G.S. (2011) Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice*, 51, 91–98. http://dx.doi.org/10.1016/j.scijus.2011.03.002

Morrison, G.S. (2012) The likelihood-ratio framework and forensic evidence in court: A response to R v T. *International Journal of Evidence and Proof*, 16, 1–29. http://dx.doi.org/10.1350/ijep.2012.16.1.390

Morrison, G.S. (2013). Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, 45, 173–197. http://dx.doi.org/10.1080/00450618.2012.733025

Morrison, G.S. (2014) Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison. *Science & Justice*, 54, 245–256. http://dx.doi.org/10.1016/j.scijus.2013.07.004

Morrison, G.S. (2017) What should a forensic practitioner's likelihood ratio be? II. *Science & Justice*, 57, 472–476. http://dx.doi.org/10.1016/j.scijus.2017.08.004

Morrison, G.S. (2018a) Admissibility of forensic voice comparison testimony in England and Wales. *Criminal Law Review*, 2018(1), 20–33. (Available from http://geoff-morrison.net/#Admissibility_EW_2018)

Morrison, G.S. (2018b) The impact in forensic voice comparison of lack of calibration and of mismatched conditions between the known-speaker recording and the relevant-population sample recordings. *Forensic Science International*, 283, e1–e7. http://dx.doi.org/10.1016/j.forsciint.2017.12.024

Morrison, G.S., Ballantyne, K., Geoghegan, P.H. (2018) A response to Marquis et al (2017) What is the error margin of your signature analysis? *Forensic Science International*, 287, e11–e12. https://doi.org/10.1016/j.forsciint.2018.03.009

Morrison, G.S., Enzinger, E. (2018) Score based procedures for the calculation of forensic likelihood ratios – Scores should take account of both similarity and typicality. *Science & Justice*, 58, 47–58. http://dx.doi.org/10.1016/j.scijus.2017.06.005

Morrison, G.S., Enzinger, E., Hughes, V., Jessen, M., Meuwly, D., Neumann, C., Planting, S., Thompson, W.C., van der Vloed, D., Ypma, R.J.F., Zhang, C., Anonymous, A., Anonymous, B. (2021) Consensus on validation of forensic voice comparison. *Science & Justice*, 61, 229–309. https://doi.org/10.1016/j.scijus.2021.02.002

Morrison, G.S., Enzinger, E., Ramos, D., González-Rodríguez, J., Lozano-Díez, A. (2020) Statistical models in forensic voice comparison. In *Handbook of Forensic Statistics* (eds D. Banks, K. Kafadar, D.H. Kaye, M. Tackett), pp. 451–497. Boca Raton, FL: CRC. https://doi.org/10.1201/9780367527709

Morrison, G.S., Kaye, D.H., Balding, D.J., Taylor, D., Dawid, P., Aitken, C.G.G., Gittelson, S., Zadora, G., Robertson, B., Willis, S.M., Pope, S., Neil, M., Martire, K.A., Hepler, A., Gill, R.D., Jamieson, A., de Zoete, J., Ostrum, R.B., Caliebe, A. (2017) A comment on the PCAST report: Skip the "match"/"non-match" stage. *Forensic Science International*, 272, e7–e9. http://dx.doi.org/10.1016/j.forsciint.2016.10.018

Morrison, G.S., Neumann, C., Geoghegan, P.H. (2020) Vacuous standards – subversion of the OSAC standards-development process. *Forensic Science International: Synergy*, 2, 206–209. https://doi.org/10.1016/j.fsisyn.2020.06.005

Morrison, G.S., Neumann, C., Geoghegan, P.H., Edmond, G., Grant, T., Ostrum, R.B., Roberts, P., Saks, M., Syndercombe Court, D., Thompson, W.C., Zabell, S. (2021) Reply to Response to Vacuous standards – subversion of the OSAC standards-development process. *Forensic Science International: Synergy*, 3, article 100149. https://doi.org/10.1016/j.fsisyn.2021.100149

Morrison, G.S., Sahito, F.H., Jardine, G., Djokic, D., Clavet, S., Berghs, S., Goemans Dorny, C. (2016) INTERPOL survey of the use of speaker identification by law enforcement agencies. *Forensic Science International*, 263, 92–100. http://dx.doi.org/10.1016/j.forsciint.2016.03.044

Morrison, G.S., Stoel, R.D. (2014) Forensic strength of evidence statements should preferably be likelihood ratios calculated using relevant data, quantitative measurements, and statistical models – a response to Lennard (2013) Fingerprint identification: How far have we come? *Australian Journal of Forensic Sciences*, 46, 282–292. http://dx.doi.org/10.1080/00450618.2013.833648

Morrison, G.S., Thompson, W.C. (2017) Assessing the admissibility of a new generation of forensic voice comparison testimony. *Columbia Science and Technology Law Review*, 18, 326–434. https://doi.org/10.7916/stlr.v18i2.4022

National Research Council of the National Academies (2009) *Strengthening Forensic Science in the United States: A Path Forward*. Washington, DC: National Academies Press. https://doi.org/10.17226/12589

Neumann, C., Ausdemore, M. (2020) Defence against the modern arts: The curse of statistics – Part II: 'Score-based likelihood ratios'. *Law, Probability and Risk*, 19, 21–42. http://dx.doi.org/10.1093/lpr/mgaa006

Neumann, C., Hendricks, J., Ausdemore, M. (2020) Statistical support for conclusions in fingerprint examinations. In *Handbook of Forensic Statistics* (eds D. Banks, K. Kafadar, D.H. Kaye, M. Tackett), pp. 277–324. Boca Raton, FL: CRC. https://doi.org/10.1201/9780367527709

Ommen, D.M., Saunders, C.P. (2021) A problem in forensic science highlighting the differences between the Bayes factor and likelihood ratio. *Statistical Science*, 36, 344–359. https://doi.org/10.1214/20-STS805

President's Council of Advisors on Science and Technology (2016) Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. (Available from https://obamawhitehouse.archives.gov/administration/eop/ostp/pcast/docsreports/)

Prince, S.J.D., Elder, J.H. (2007). Probabilistic linear discriminant analysis for inferences about identity. In Proceedings of the IEEE 11th International Conference on Computer Vision, pp. 1–8. https://doi.org/10.1109/ICCV.2007.4409052

Ramos, D., Meuwly, D., Haraksim, R., Berger, C.E.H. (2020) Validation of forensic automatic likelihood ratio methods. In *Handbook of Forensic Statistics* (eds D. Banks, K. Kafadar, D.H. Kaye, M. Tackett), pp. 143–163. Boca Raton, FL: CRC. https://doi.org/10.1201/9780367527709

Redmayne, M., Roberts, P., Aitken, C.G.G., Jackson, G. (2011). Forensic science evidence in question. *Criminal Law Review*, 2011(5), 347–356. (Available from https://rke.abertay.ac.uk/en/publications/forensic-science-evidence-in-question)

Risinger, D.M. (2013) Reservations about likelihood ratios (and some other aspects of forensic 'Bayesianism'). *Law, Probability and Risk*, 12, 63–73, http://dx.doi.org/10.1093/lpr/mgs011

Roux, C., Weyermann, C. (2021) From research integrity to research relevance to advance forensic science. *Forensic Sciences Research*. https://doi.org/10.1080/20961790.2021.1977480

Roux, C., Willis, S., Weyermann, C. (2021) Shifting forensic science focus from means to purpose: A path forward for the discipline? *Science & Justice*, 61, 678–686. https://doi.org/10.1016/j.scijus.2021.08.005

Saks, M.J., Koehler, J.J. (2005) The coming paradigm shift in forensic identification science. *Science*, 309, 892–895. https://doi.org/10.1126/science.1111565

Saks, M.J., Koehler, J.J. (2008) The individualization fallacy in forensic science. *Vanderbilt Law Review*, 61 199–219. (Available from https://ssrn.com/abstract=1432516)

Stoel, R.D., Berger, C.E.H., Kerkhoff, W., Mattijssen, E.J.A.T., Dror, E.I. (2015) Minimizing contextual bias in forensic casework. In *Forensic Science and the Administration of Justice: Critical Issues and Directions* (eds K.J. Strom, M.J. Hickman), pp. 67–86. Thousand Oaks CA: Sage. http://dx.doi.org/10.4135/9781483368740.n5

Swofford, H., Champod, C. (2021) Implementation of algorithms in pattern & impression evidence: A responsible and practical roadmap. *Forensic Science International: Synergy*, 3, article 100142. https://doi.org/10.1016/j.fsisyn.2021.100142

Swofford, H., Cole, S., King, V. (2021) Mt. Everest – we are going to lose many: A survey of fingerprint examiners' attitudes towards probabilistic reporting. *Law, Probability and Risk*, 19, 255–291. https://doi.org/10.1093/lpr/mgab003

Thompson, W.C. (2012) Discussion paper: Hard cases make bad law – reactions to R v T. *Law, Probability and Risk*, 11, 347–359. https://doi.org/10.1093/lpr/mgs020

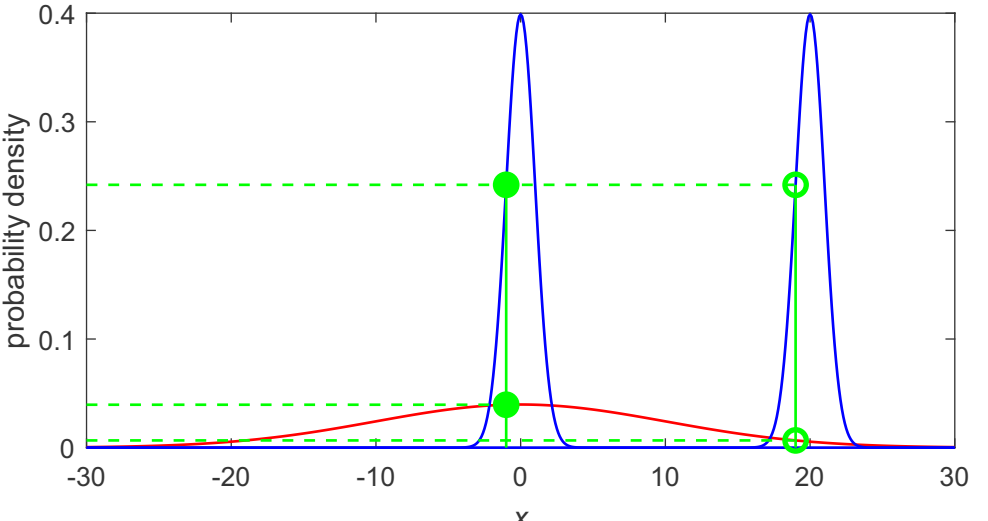
1151 Thompson, W.C., Black J., Jain A., Kadane J. (2017) Forensic Science Assessments:
1152 A Quality and Gap Analysis- Latent Fingerprint Examination. Washington, DC:
1153 American Association for the Advancement of Science.
1154 https://doi.org/10.1126/srhrl.aag2874

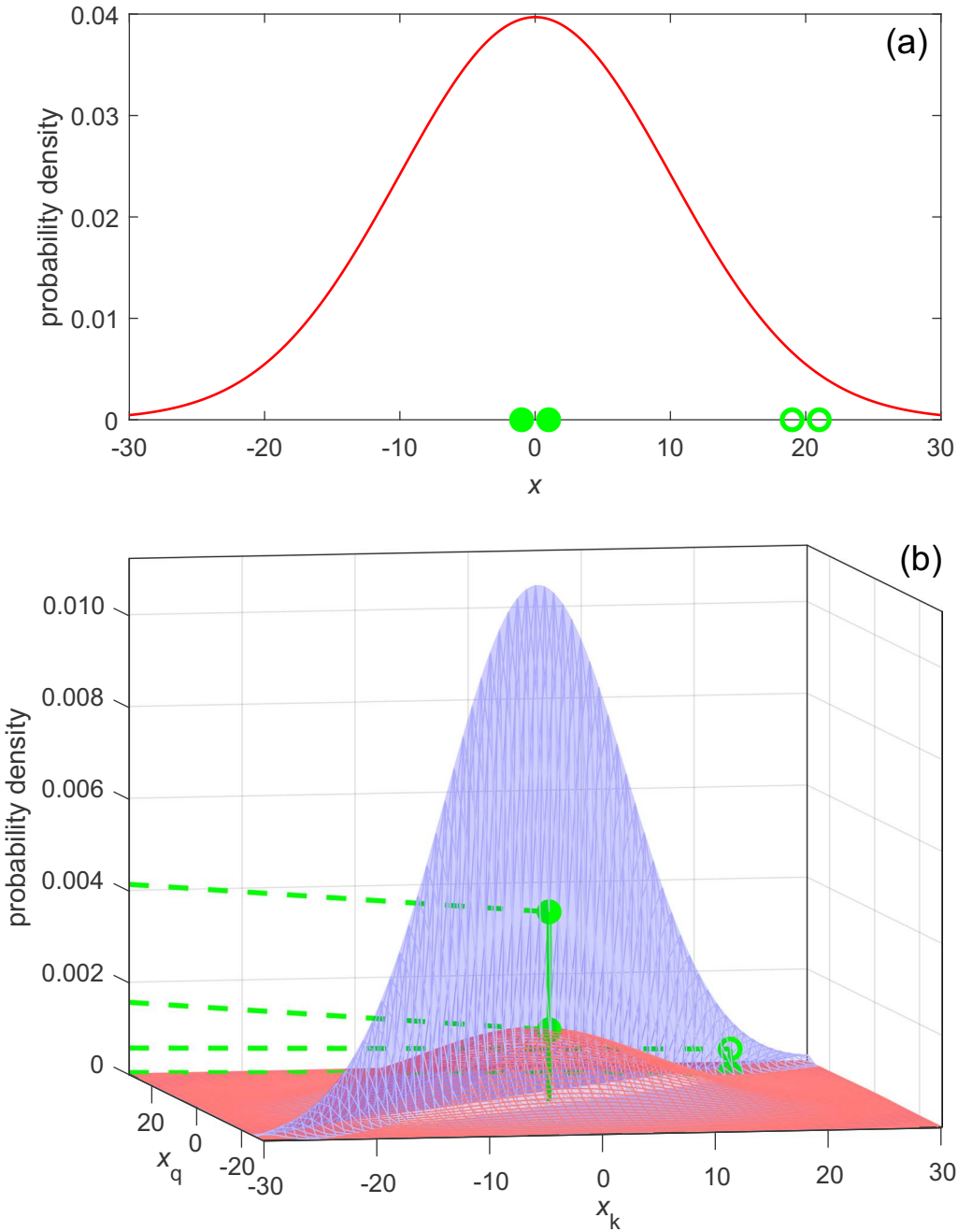
1155 Thompson, W.C., Schumann, E.L. (1987) Interpretation of statistical evidence in
1156 criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law*
1157 *and Human Behavior*, 11(3), 167–187. http://www.jstor.org/stable/1393631

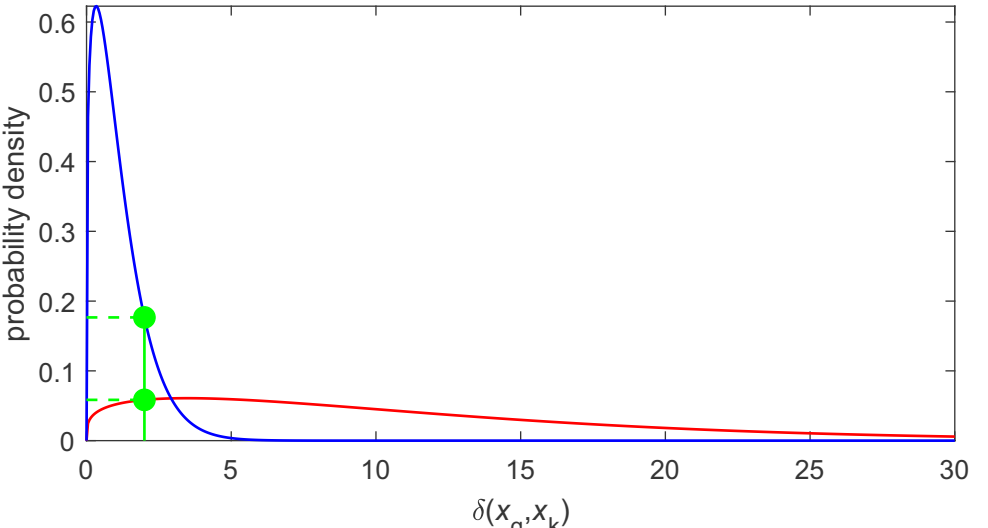
1158 Villalba, J., Chen, N., Snyder, D., García-Romero, D., McCree, A., Sell, G.,
1159 Borgstrom, J., García-Perera, L.P., Richardson, F., Dehak R., Torres-
1160 Carrasquillo, P.A., Dehak, N. (2020) State-of-the-art speaker recognition with
1161 neural network embeddings in NIST SRE18 and Speakers in the Wild
1162 evaluations. *Computer Speech & Language*, 60, article 101026.
1163 https://doi.org/10.1016/j.csl.2019.101026

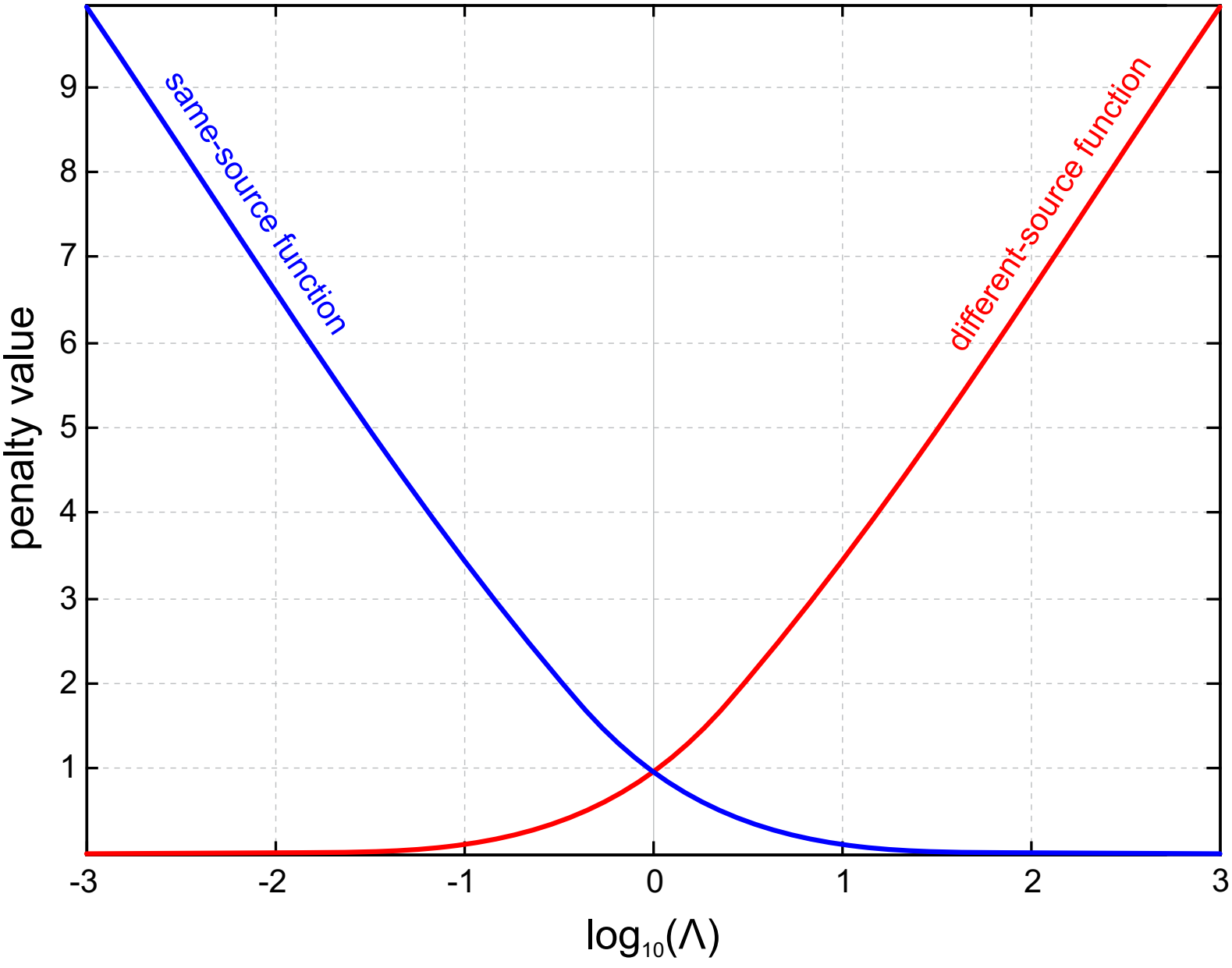
1164 Weber, P., Enzinger, E., Labrador-Serrano, B., Lozano-Díez, A., Ramos, D.,
1165 González-Rodríguez, J., Morrison, G.S. (2022 in press) Forensic voice
1166 comparison – Human-supervised-automatic approach. In Encyclopedia of
1167 Forensic Sciences (3rd ed). Elsevier.

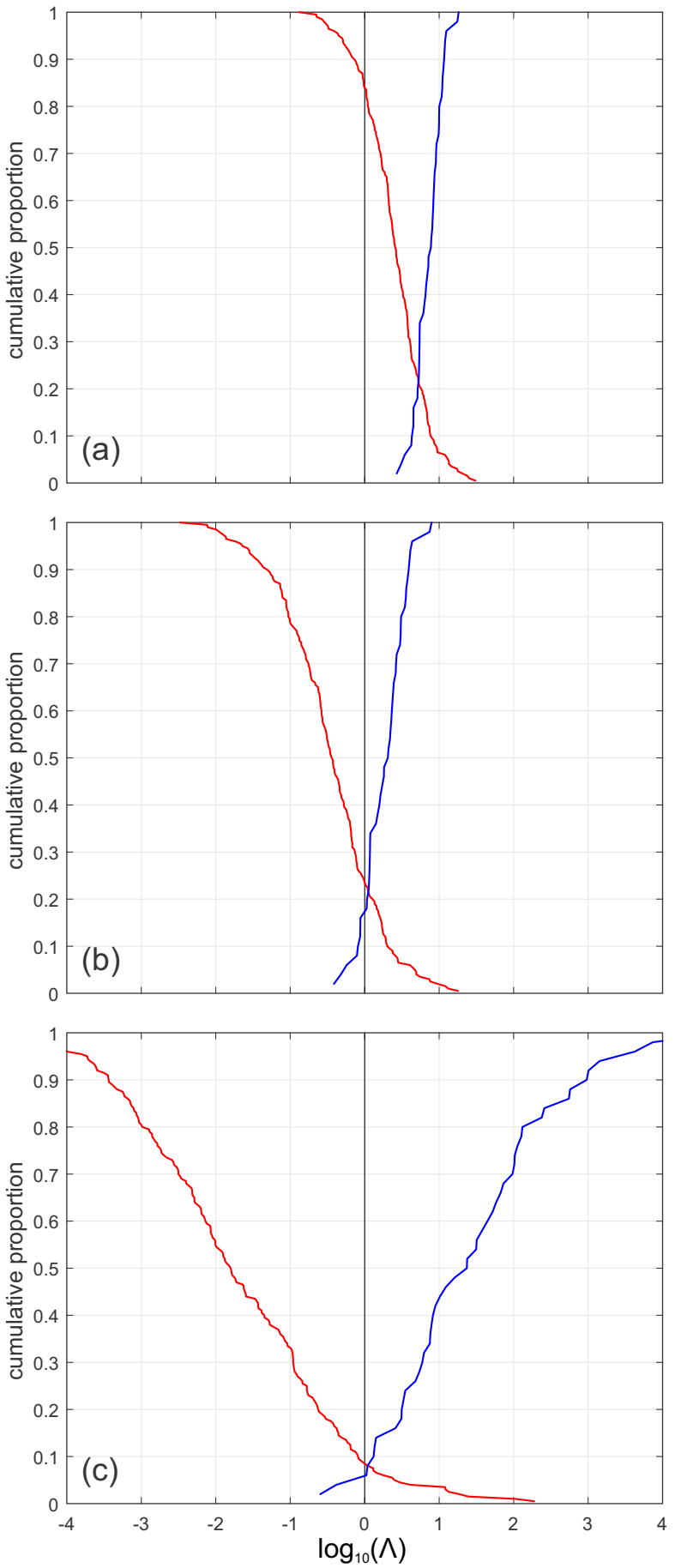
1168 Willis, S.M., McKenna, L., McDermott, S., O'Donell, G., Barrett, A., Rasmusson,
1169 A., Nordgaard, A., Berger, C.E.H., Sjerps, M.J., Lucena-Molina, J.J., Zadora,
1170 G., Aitken, C.G.G., Lunt, L., Champod, C., Biedermann, A., Hicks, T.N.,
1171 Taroni, F. (2015) ENFSI Guideline for Evaluative Reporting in Forensic
1172 Science. (Available from http://enfsi.eu/wp-
1173 content/uploads/2016/09/m1_guideline.pdf)


0.4
0.3
0.2
0.1
0
probability density
-30
-20
-10
0
10
20
30
x


(a)
probability density
0.04
0.03
0.02
0.01
0
-30
-20
-10
0
10
20
30
x
(b)
probability density
0.010
0.008
0.006
0.004
0.002
0
20
0
-20
x_q
-30
-20
-10
0
10
20
30
x_k


probability density
0.6
0.5
0.4
0.3
0.2
0.1
0
0
5
10
15
20
25
30
δ(x_q,x_k)


penalty value
9
8
7
6
5
4
3
2
1
-3
-2
-1
0
1
2
3
log10(Λ)
same-source function
different-source function


cumulative proportion
(a)
(b)
(c)
log10(Λ)