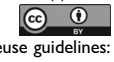
Article

THE INTERNATIONAL JOURNAL of EVIDENCE & PROOF

Why the post-identification era is long overdue: Commentary on the current controversy over forensic feature comparison as applied to forensic firearms examination

The International Journal of Evidence & Proof
2025, Vol. 29(2) 140–160


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/13657127241278069
journals.sagepub.com/home/epj


Alex Biedermann and **Christophe Champod**
Faculty of Law, Criminal Justice and Public Administration, School of Criminal Justice, University of Lausanne, Lausanne, Switzerland

Abstract

In this commentary, we critically review recurring arguments for and against the discipline of forensic feature comparison as applied to firearms examination from various commentators within and outside forensic science. One of the mainstream criticisms that we address, among others, is that the field cannot demonstrate sufficient proficiency and robustness based on empirical (i.e., black-box) studies. While the lack of empirically demonstrated examiner proficiency is a valid concern and a powerful concept in the short term (e.g., in admissibility proceedings), many critics reduce their discussion of forensic feature comparison solely to the need to measure and demonstrate proficiency through error rates. However, the exclusive focus on aggregate expert performance metrics, here referred to as *examiner diagnosticism*, remains a surface-level perspective. It provides an incomplete account of the field because these metrics do not represent—but are often confused with—the notion of the evidentiary value of findings, i.e., observations made on examined items in individual cases. We argue that examiner diagnosticism should be contrasted and complemented with the notion of *feature selectivity*, i.e., the diagnostic capacity of observed marks and features on examined items. We argue that forensic scientists should report and be probed on their ability to quantify feature selectivity (i.e., the probative value of findings). By ceasing to express source attribution opinions (identification/individualisation), which are now widely exposed as unscientific, the forensic feature comparison disciplines could move further into the long-awaited post-identification era pioneered by other fields such as forensic genetics.

Corresponding author:
Alex Biedermann, Faculty of Law, Criminal Justice and Public Administration, School of Criminal Justice, University of Lausanne, Lausanne, Switzerland
Email: Alex.Biedermann@unil.ch

Keywords
Aggregate performance measures, error rates, forensic feature comparison, identification, interpretation, value of evidence

Introduction

The debate over the scientific status of the field of forensic feature comparison as applied to firearms examination, especially in the United States, has recently grown more heated. The positions espoused by the various participants in this discussion range widely. Consider a brief, non-exhaustive illustration of the landscape. In the forefront are the proponents of the Association of Firearm and Tool Mark Examiners (AFTE) Theory of Identification (Committee for the Advancement of the Science of Firearm & Toolmark Identification, 2011). Many of them are practitioners who claim to be able to make source attribution determinations. In the context here, a source attribution determination, also called identification or individualisation,[1] is the assertion that a particular weapon (or part of it, such as a barrel or firing pin) was used to fire a particular item of ammunition, such as a bullet or a cartridge case. Academic circles, on the other hand, have traditionally rejected this position. They dispute the feasibility of categorical source attribution conclusions and object to their unscientific nature (e.g., Saks and Koehler, 2005; Schwartz, 2005; Stoney, 1991). With respect to legal practice, it is worth noting that several courts in the United States have recently become the focus of attention as they have carefully examined the matter through evidentiary hearings and affidavits from various experts, including experts *other* than forensic firearms examiners.[2] These so-called 'anti-expert experts' are 'classically trained research scientists who could verify claims made by firearms examiners and explain basic principles and methods of science' (Faigman et al., 2022).

In this article, we set out to review some of the key assertions made by the various participants in the debate over forensic feature comparison. We seek to draw broader and more nuanced conclusions than those that emerge from mainstream arguments at the level of individual position statements, often tailored exclusively for or against the field.

This commentary is structured as follows. The next section begins with a brief overview of the major problems that characterise the field of forensic feature comparison and how critics outside of forensic science portray the field. We then present and discuss points of agreement and disagreement with these critical positions and put our contrasting viewpoints into perspective. Next, we propose a series of changes that we believe are necessary for the field of forensic feature comparison, both in research and practice, to achieve a more scientifically defensible foundation. Finally, we present a discussion and conclusions.

How critics portray forensic feature comparison as applied to firearms examination

Firearms examination and identification based on traces left on fired bullets and cartridge cases has a long and contentious history. In the 1960s, Berkeley Professor Paul Kirk wrote: 'Many persons can identify the particular weapon that fired a bullet, but few if any can state a single fundamental principle of identification of firearms' (Kirk, 1963). It is no exaggeration to say that this is still the case for many

1. For the purpose of this commentary we treat identification and individualisation, as is often done in informal conversation, as synonyms. However, we recognise that, strictly speaking, identification is merely classification, i.e., the assignment of an object to a class of objects of a certain kind (Kirk, 1963).
2. See e.g., *Kobina Ebo Abruquah* v. *State of Maryland*, No. 10, September Term, 2022 (Opinion by Fader, C.J.).

examiners today, because commonly invoked accounts, such as the AFTE Theory of Identification (AFTETI) (Committee for the Advancement of the Science of Firearm & Toolmark Identification, 2011), do not provide the requisite foundation. The AFTETI is merely *descriptive* of the *judgements* made by examiners when they say that they reached an identification. The theory does not and cannot warrant the knowledge claim that identification conclusions suggest, i.e., the assertion of excluding all other potential sources. It is not surprising, then, that categorical conclusions offered by forensic feature comparison analysts have been and continue to be viewed with suspicion, especially by defence lawyers and academic critics. In the forthcoming sections, we will take a closer look at and discuss these and other principal lines of criticism of forensic feature comparison.

In a recent opinion paper entitled 'The field of firearms forensics is flawed', published in the *Scientific American*, Faigman et al. (2022) start with the observation that 'the matching of bullets is subjective' and that 'few studies of firearms exist and those that do indicate that examiners cannot reliably determine whether bullets or cartridges were fired by a particular gun'. We refer to this paper here because it represents, in many ways, an essential and recurring point of criticism of forensic feature comparison as applied to firearms examination.

According to Faigman et al. (2022), firearms examiners who present themselves in courts as experts are (merely) *practitioners* who apply a forensic technique, i.e., feature comparison. One should distinguish these examiners, still according to the same authors, from research scientists who are 'professionally trained in experimental design, statistics and the scientific method' (Faigman et al., 2022). As a means to illustrate this distinction, the authors draw an analogy to the practising nurse, capable of administering a COVID-19 vaccine but lacking the expertise of the 'research scientists who understand how it [the vaccine] was created and tested' (Faigman et al., 2022). In essence, the critical argument here, as we understand it, is that 'only research scientists have the wherewithal to counter the claims of practitioner-experts' and that such research scientists can, therefore, take the role of so-called 'anti-expert experts' (Faigman et al., 2022). That is, *only* the latter have the expertise to scrutinising empirical studies regarding the performance of firearms examiners and 'explain why these studies are flawed' (Faigman et al., 2022). Underlying the above statements are two points. The first point is that forensic feature-examiners lack a scientific attitude and competence in research design, data analysis and interpretation; hence courts need assistance from experts who possess that expertise. We will not discuss this allegation of flawed scientific background of forensic feature comparison experts, individually or as a community, because this would require specific examples beyond the scope of this article. The second point is that the field's trustworthiness hinges primarily on examiner performance and empirical evidence supporting it, as advocated prominently in the report of the PCAST (2016) and in other reviews of forensic science in the past. In the remainder of this article, we will examine and comment on this empirical perspective in terms of its pertinence, foundations and completeness for providing the ground for a fair assessment of the potential and limitations of forensic feature comparison.

Discussion of the critics' position

Points of agreement

From a broad perspective, Faigman et al. (2022) adequately assess the currently most widely used method of comparing traces on elements of ammunition and reporting the results of such comparisons. Indeed, most practising examiners reach conclusions by invoking a vaguely defined *personal assessment* of 'sufficient agreement' (Committee for the Advancement of the Science of Firearm & Toolmark Identification, 2011) between compared features. As such, the critics provide a descriptive account of the scientific status of the work provided by practitioners of forensic feature comparison.

The authors are also correct in stating that there are studies showing that a certain proportion of the surveyed examiners made errors. However, there is considerable debate about the definition of error(s)

other than in obvious situations, such as when an examiner makes a source attribution conclusion when the items being compared actually come from different sources (i.e., a so-called false positive). There is also considerable debate about how to summarise count data from examiner performance studies in order to produce performance metrics such as error rates. For reasons that will become clear in due course, we will not go into such quantitative details of existing examiner performance studies and the conclusions, if any, that they support.[3]

In the remainder of this article, we will instead focus on critically examining the premises and conceptual architecture that underlie the critics' account of the discipline of feature comparison. In this context, we will also examine the limitations of the mainstream empirical studies on which the critics rely and which seem to influence the positions they take in the current controversy.

Points of disagreement

Preliminaries: feature selectivity vs. examiner diagnosticity. In this section, we will argue that the critics, while defending a largely valid position in terms of describing where things currently stand in practice, generally rely on what we will call a *surface-level perspective*. We have chosen the term *surface-level perspective* because it does not provide clear guidance as to the deeper scientific direction in which forensic feature comparison research should be moving.[4] Specifically, we will argue that the widely discussed call for empirical *examiner* testing,[5] i.e., ground truth or 'black-box' (PCAST, 2016: 5) testing, tends to lead to misunderstandings of the notion of probative value and flawed uses of terminology involving the term 'decision' in much of the contemporary forensic science literature. In recent years, this literature has grown to such an extent, and has even spawned entire research programmes, that a critical discourse on the foundations of the underlying perspective is in order. The fundamental problem we have in mind is the distinction between *feature selectivity* on the one hand and *examiner diagnosticity* on the other. These are two distinct concepts, yet there is persistent confusion between them. Often commentators conflate them, using one as a proxy for the other. We explain the two terms below.

The concept of *feature selectivity* refers to the characteristic(s) or feature(s) and measures thereof on examined items. The forensic question of interest here is the extent to which features can help us reduce the population of potential sources. The answer to this question depends on how *selective* the features are.[6] The PCAST report, for example, refers to this as the 'relative rarity or commonality of the particular marks or features examined' (PCAST, 2016: 4, 34). In this sense, selectivity is a function of the rarity of the features; the rarer the features, the more selective they are (i.e., the greater their ability to help reduce the population of potential sources).[7]

3. See, e.g., Swofford et al. (2024) for an overview and Biedermann and Kotsoglou (2021) for a discussion of some of the irrational aspects that characterise contemporary discussions of error rate studies.
4. We acknowledge that the term *surface-level perspective* may be too dismissive from the point of view of the consumer of expert evidence, as an expert's source attribution may appear to be closer to the ultimate issue than other reporting formats. However, it is important to remember that this apparent proximity comes at the cost of the conceptual and practical limitations of source attribution conclusions.
5. Note that our focus here is only on the empirical testing of *examiners*, not on empirical testing in general (e.g., the validation of computational procedures that output value of evidence expressions in areas such as forensic voice comparison or fingerprint examination).
6. In addition, features should ideally be reproducible (i.e., stable over time), at least to some extent, because otherwise, the idea of helping with inference of shared source, based on observed similarities between compared items, would be compromised. For further discussion of the notion of reproducibility see also Champod et al. (2016).
7. For example, the information that a stain is human blood has no selectivity as to who the source is because all humans have human blood in their bodies. However, the information that the stain is human blood of a particular group, such as group A, has some selectivity because people have different blood groups.

As an aside, we add that the focus on feature selectivity may be characterised as the *internal view* (Biedermann, 2022). That is, we emphasise the viewpoint of the examiner who observes features and seeks to elucidate their probative value (selectivity). The usefulness of introducing the notion of internal view(point) will become apparent in due course, when we will contrast it with the notion of external view.

Consider now the notion of *examiner diagnosticity*. This notion draws the attention *exclusively* to the examiner's conclusion, screened off from the observed features. We call this the *external view* (Biedermann, 2022) because it is concerned with viewing examiners from an outside position as they perform a specific task (here: feature comparison). That is, unlike feature selectivity, the focus is *not* on the informative value of the actual features of the compared items in the case at hand. Instead, one considers only the examiner's response, and the question asked is how *diagnostic* this response is with respect to the ground truth (here: whether or not the items being compared come from the same source).

Currently, the prime tool to obtain information about examiner diagnosticity is the black-box study, prominently featured in the PCAST report: 'For subjective feature comparison methods, appropriately designed black-box studies are required, in which many examiners render decisions about many independent tests (typically, involving "questioned" samples and one or more "known" samples) and the error rates are determined' (PCAST, 2016: 46). Though there is much discussion about how to design such studies and summarise resulting data adequately (Swofford et al., 2024), we can state the central point for our purposes here as follows: the result of such black-box studies will—at best—lead to *aggregate* (i.e., overall) measures of performance of an individual examiner or a group of examiners under the particular controlled conditions of the study at hand. This broad focus entails a series of convoluted matters that we address in the next section.[8]

The fundamental problem associated with examiner diagnosticism. Given the distinctions introduced above, we are ready to explain the fundamental problem associated with examiner diagnosticity that affects a considerable part of contemporary forensic science literature and discussions among practitioners. The problem arises in the context of discourses about the probative value of expert evidence as given by an examiner's conclusion. Specifically, some discussants (e.g., Guyll et al., 2023; Smith and Neal, 2021) argue that what one should infer from an expert's conclusion regarding a feature comparison is the positive predictive value (PPV),[9] analogously to what is done with conventional diagnostic tests.[10] Below, we explain why this analogy is misconceived and why it leads to a series of conundrums.

Technically speaking, one can, of course, compute a PPV based on aggregate expert performance metrics, but in practice, the result will be close to meaningless. The reason for this is the flawed analogy between traditional (e.g., medical) diagnostic tests and forensic examiners, who are assumed to function as a black box. In diagnostic tests, the item to which a test is applied contains (or does not contain) a well-defined target substance.[11] The task consists in detecting that substance. Stated otherwise, one inquires about the test's ability to detect the target. The items examined during medical validation

8. See also Swofford et al. (2024) for an argument that summaries of validation study data lead to a loss of information for coherent decision making by recipients of expert information.
9. The positive predictive value (PPV), also sometimes called the Predictive Value Positive (PVP) (e.g., Gastwirth, 1987; Kaye, 1987), is the probability that a given examined person or item has the condition of interest (e.g., a disease) *given* that the test classified the person or item as positive.
10. See Rosenblum et al. (2024) for a critical discussion of the misuse of this framework in a real case.
11. Note that this is a simplification in the sense that the ultimate purpose may not be limited to the detection of a particular substance. Instead the goal may be to draw a conclusion about a higher-level proposition (e.g., the presence or absence of a particular disease). In this sense, a target substance is itself an indicator of a particular condition (or disease). A test may also have multiple targets. In the context of medical or psychological tests, the focus may be on the broader concept of 'symptom' rather than on the notion of target substance (e.g., Kaye, 1987).

studies, e.g., good-quality human blood samples, are of the *same kind* (in terms of quality and quantity) as those encountered when one deploys the test in operational practice. Due to these situational characteristics, one can, for example, consider the sensitivity,[12] as determined in the test's validation study, to be informative about the probability of the test turning out to be positive in an operational application where an item contains the target substance in a detectable quantity or concentration.

Forensic feature comparison, however, has (almost) nothing to do with this. When an examiner concludes 'identification' (analogous to a diagnostic test indicating 'positive'), such an utterance is strictly *uninformative* about the features on which the examiner has based the conclusion.[13] The compared features could be virtually anything in terms of quality and quantity of striation marks (e.g., on fired bullets) judged to be (sufficiently) similar or corresponding. More specifically, there is no standard (i.e., predefined) and recurrent target characteristic (or combination of characteristics) in forensic feature comparison, akin to a target substance in medical diagnosis. Note that the point here is not a problem of unstable (i.e., non-reproducible) occurrence of target features. The point is that, at a sufficiently deep level of observational resolution, each source has its own feature configuration. It is this *a priori* indeterminacy of feature configurations that undermines the usefulness of the analogy with the classic diagnostic testing perspective.[14]

In light of the above, the strict black-box treatment of forensic feature comparison experts as diagnostic test devices, i.e., calculating the PPV based solely on the expert's utterance of an identification (or other categorical conclusion), leads to three interrelated contortions:

1. the expert's utterance is taken as evidence, rather than the actual features of the items being compared,
2. the *same* overall conclusion (here: PPV) would be drawn[15] from an expert's declared identification regardless of the actual feature configuration (e.g., its quality),[16]
3. the overall diagnostic performance of the examiner would be conflated with (and used as a surrogate or proxy for) the selectivity of the features of the examined items in the case at hand.[17]

However, the contortions do not end here. The analogy with diagnostic testing is also flawed at a deeper level. Note that (medical) diagnosis is simply a form of classification. It involves recognising items as belonging (or not) to a particular class which consists of multiple members that share the same label (such as having a particular disease) (e.g., Gastwirth, 1987). Class attribution is based on detecting a particular, agreed-upon target substance. However, this understanding cannot be applied to forensic source

12. The sensitivity of a test is the probability that a test turns out to be 'positive' (i.e., indicates the presence of a target substance) *if* the tested item or person indeed contains the target substance (or, has a particular condition).
13. See also 'A graphical illustration of the flawed analogy between conventional diagnostic testing and black-box testing of forensic examiners' for a more detailed explanation.
14. See also Imwinkelried (2020) for a discussion of additional aspects, such as the notion of range of validation, regarding the discrepancy between the characteristics of the case at hand and the experimental conditions used in the validation study.
15. Of course, the PPV depends not only on the sensitivity but also on the specificity (i.e., the probability of obtaining a negative result in the absence of the target substance or condition) and the initial (or prior) probability (e.g., Gastwirth, 1987). In medical screening applications, it is common to refer to the notion of base rate (e.g., the prevalence of a disease in a population of interest) rather than the notion of prior probability. However, we will avoid the term base rate here because of its connotation with frequentist statistics and the confusion it can cause (for a discussion, see e.g., Biedermann and Vuille (2018)). Our focus here is on specific cases (i.e., unique historical events), not on an item (or person) randomly drawn from a population of interest, disconnected from the context of the case at hand.
16. This amounts to nothing less than ignoring the actual feature configuration, especially its information content.
17. There is a way to treat both the potential for examiner error *and* the rarity of features coherently in a single cascaded inference model (Thompson et al., 2003). See also Schum (1994) for yet more detailed analyses of various attributes that characterise the credibility of human sources of information. The details of these formal evaluation frameworks are beyond the scope of this commentary.

attribution (or identification). In forensic source attribution problems, each potential source represents its own class. Source attribution is thus a multi-class inference problem, where each class consists of a single source that is distinct from all others in terms of its own distinctive feature configuration.

Of course, one can assume, as some approaches do (e.g., Smith and Neal, 2021), that identification is a binary classification problem where the two classes are 'same-source' and 'different-source'. That is, the examiner tries to assert whether the given *pair* of compared items belongs to the category of same-source or different-source pairs. However, as noted above, there is no class-wide target feature in this account. The diagnostic inference framework can only be applied here by bypassing the actual features and using the examiner's mere utterance as a crude 'indicator' of the main propositions of interest (i.e., same- vs. different-source). We see, then, that forensic source inference is much more complex than the binary analogy to classical diagnostic testing suggests, and than proponents of this approach would have us believe.

The bottom line is that there is considerable confusion about concepts about which there should reign clarity. This confusion can lead discussants to mistakenly believe, for example, that they are talking about the evidentiary value of *specific* feature comparisons in a particular case, when in fact they are only talking about the *general* diagnosticity of expert utterances. In other words, general characterisations of expert performance *in the aggregate case*, as measured by black-box studies, are not a measure, even a proxy, for the value of feature configurations observed on compared items in individual cases.[18]

As an interim conclusion, we note that the critics' focus on examiner performance and its assessment through empirical studies, while a relevant *general* consideration,[19] represents only a limited aspect of the discipline. It is a surface-level perspective because it does not address the potential probative value associated with case-specific findings (i.e., features observed on compared items) and how this information might shape the examiner's opinion.

A graphical illustration of the flawed analogy between conventional diagnostic testing and black-box testing of forensic examiners

Consider a simple graphical model to represent the logical structure of conventional diagnostic testing. As shown in Figure 1(i), the result of a test (represented by the node T) depends on the presence (or absence) of a target substance (node S) that the test is designed to detect. In turn, the presence of the target substance S in an examined item depends on the condition of the person being examined, i.e., whether or not the person has a particular disease (represented by the node D). In the graphical language used here, variables are represented by nodes, and the statement that a variable A depends on (or is influenced by) another variable B is expressed by a directed arc pointing from the latter to the former variable (e.g., Taroni et al., 2014).

Figure 1(ii) represents a model for forensic inference of source (Taroni et al., 2004; Thompson et al., 2003). The node R represents a scientist's report of observed corresponding features between an item of unknown source and an item of known source. Such a report depends on whether the compared items actually have corresponding features (proposition represented by the node F), as reported by the examiner. In turn, whether or not the compared items share corresponding features, proposition F, depends on whether or not they come from the same source (represented by the node H). The dashed lines in

18. See also Kaye (1987) for further discussion of why the PPV (or PVP) is not an appropriate concept for expressing probative value.

19. One way to acknowledge the value of examiner diagnosticism and forensic black box testing is to see these perspectives as examples of 'framework evidence' (Faigman et al., 2014). In turn, the notion of feature selectivity would lean more towards case-tailored 'diagnostic evidence' (Faigman et al., 2014).

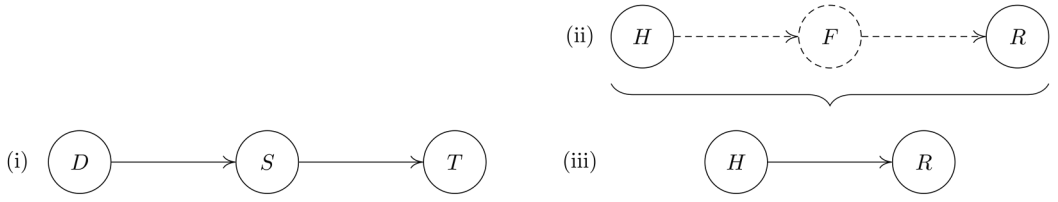


Figure 1. Graphical models for (i) conventional diagnostic testing, (ii) forensic source inference and (iii) forensic black-box testing. The definition of the variables is given in the text.

Figure 1(ii) indicate the elements that are ignored by forensic black-box testing, shown in Figure 1(iii). Forensic black-box testing only records examiners' reports (R) for comparison pairs from the same source or from different sources (H). This structural assumption is problematic, as further discussed below.

When populated with data from a black-box study, the model in Figure 1(iii) can at best help answer inductive questions of the following type (aligned with conventional diagnostic reasoning): 'What is the probability that an unspecified pair of compared items from a population of the type studied in the validation/black-box study came from the same source, *given* the examiner's report of observed similarities and differences (or, alternatively, a reported "identification")?' Strictly speaking, this is a conditional probability statement about a target proposition (here the proposition 'the compared items come from the same source'), similar to a PVP in diagnostic testing. However, it is disconnected from the specific case at hand for two reasons. First, because it does not take into account any other evidence that might have a bearing on the case at hand. Second, even if other evidence had been properly accounted for (by adjusting the prior probability), the statement would still not reflect the probative value of the features observed during comparison, because they are not part of the considerations. Specifically, the observed feature configuration in the case at hand may be anything from completely uninformative to highly selective.

Thus, sweeping away feature selectivity and reducing considerations to the diagnosticity of examiners in the *aggregate case* is prone to misrepresent, either by overvaluing or undervaluing, the informative value associated with the observed features (Biedermann, 2022). Overall, this problem is an example of what has been discussed elsewhere in the legal literature as the problem of 'Group to Individual (G2i) Inference' (Faigman et al., 2014).

Placing our critique into perspective

The points raised in the previous section may seem to contradict the position of the critics outlined at the beginning of this article and may therefore be seen as unscientific, anti-empirical or retrograde. In this section we anticipate three possible objections in order to provide more context for our perspective.

First, we stress that we are not suggesting that general performance metrics, such as error rates, are not important or useful. Aggregate performance metrics have their place as an essential piece of information in discussions about the admissibility of a particular type of evidence and/or a particular examiner in the requisite field. But, and we repeat, the information conveyed by aggregate performance measures—even if they can serve as an anchor for assessments in individual cases (Koehler, 2008)—remains of limited usefulness for dealing with experts and case-specific findings (i.e., comparison results) in instant cases. Even the fields of AI and machine learning have now begun to recognise the problem with aggregate performance metrics in their reporting of research findings (Burnell et al., 2023). What this means for the context of our discussion is that whenever one is concerned with assessing the informative value of a particular feature comparison in the instant case, there is no way around quantifying the selectivity of the

feature configuration that the examiner observed to be in agreement (or not) or to show similarities (or differences),[20] as the case may be.

It is also worth noting that the topic of performance measures has a strategic connotation. From the perspective of a party against which feature comparison evidence is being introduced, it makes perfect sense to argue against a discipline as a whole, without bothering to look at the details of what exactly the examiner observed (e.g., the quality and quantity of striation marks). However, this focus is insufficient to provide a comprehensive assessment of the potential and limitations of the discipline as a whole. The narrowly framed discussion centred on performance metrics is to be distinguished from broader discussions, such as ours here, which focus on the potential value of the results of feature comparisons, regardless of which party may benefit from those results. We believe that this is what a fair assessment of the discipline should aim to do.

Second, and related to the above, our critique should not be misconstrued as a rejection of ground truth testing. We fully support empirical testing using test pairs where the ground truth is known (i.e., whether or not the items come from the same source). However, the key difference with respect to black-box studies of the type advocated in the PCAST report is that we call for studies in which the value of the observed feature configuration is reported, *not* the direct opinions of examiners about source-level propositions, especially categorical source attribution conclusions. In the forensic statistics literature, such empirical performance evaluation for value of evidence assessments has seen substantial development over the past decade. Examples of relevant work in these areas are presented by Meuwly et al. (2017), Morrison et al. (2021), Ramos and Gonzalez-Rodriguez (2013) and Ramos et al. (2021). Chapter 8 of Aitken et al. (2020) provides an overview of the topic.

Third, in criticising examiner diagnosticism, we do not intend to defend unscientific state-of-the-art practices in court. As noted under 'Points of agreement', we do not deny the fact that certain feature comparison experts exhibit suboptimal performance and/or use unscientific reporting formats, neither of which should have any place in legal proceedings. Our disagreement lies in the conclusions that should be drawn from the observation of suboptimal examiner performance. Some critics categorically conclude that '[t]he field of firearms forensic is flawed' (Faigman et al., 2022). However, this is an overly general conclusion that does not adequately reflect the complexities, subtleties and potential of the field, particularly with regard to our discussion of feature selectivity.

More specifically, what critics can legitimately claim is that the practice of some, but not all, practising forensic firearms examiners is suboptimal. This difference in emphasis is a direct consequence of the critics' surface-level perspective, which focuses only on examiner diagnosticity. In contrast, our position is that there is, *in principle*, potential[21] probative value associated with intentional design features (also called class characteristics) of elements of ammunition (bullets and cartridge cases) and accidental marks thereon. Part of this potential value derives from the fact that observable features vary widely, so that even a correspondence observed at the level of a single class characteristic (e.g., the calibre) is necessarily probative, although perhaps only in a limited way. Another way of understanding this is to consider that all potential sources with manifestly different class characteristics are technically incompatible with the features of the evidential item. Some of these characteristics can be recognised even by non-specialists.

In the light of these considerations, it seems an exaggeration to suggest that the field of firearms examination *as a whole* is flawed and should be eliminated from use in the legal system. This is not what a scientific and fair assessment of the field should conclude. We should recognise that the field is not limited to examiner competence, although it depends on it, but also includes the existence of an organised

20. Here, we use descriptors that provide a realistic account of what is observed, i.e., *similarities* and dissimilarities, contrary to what other, obsolete terms, such as 'match/non-match' suggest. For further discussion, see also Morrison et al. (2017).

21. We qualify probative value here as 'potential' because it depends on the quality and completeness of the items submitted to examination.

body of knowledge about the intrinsic selectivity associated with marks and features present on examined items. This organised body of knowledge should be used to the extent that this can be done in a defensible manner. This is discussed further under 'Quantifying the probative value of comparisons of marks and features'.

One is, of course, free to ignore this potential of the field. But then one should ask, for the sake of coherence, whether one would take any comfort as a defendant in a situation where the firearms evidence favoured one's case because of clearly incompatible class characteristics. Should one still insist that the field of feature comparison is of no use *in the case at hand*? We recognise that exclusionary findings may not be a primary concern, as in practice such cases may not even make it to trial. But again, our focus here is not limited to one type of case. Our focus is on what the field of feature comparison can contribute to the legal process in principle, regardless of which party might benefit from the findings.

Sceptical readers may remain unconvinced by our emphasis on the potential probative value of features (marks) associated with elements of ammunition. They may argue that the reliable detection and use of such features inevitably depends on the competence of the examiners, since feature selectivity cannot be of practical value if the examiner cannot demonstrate competence in detecting features appropriately. We do not dispute this point, of course, but this argument confuses a scientific claim (about feature selectivity) with an operational claim (about examiner proficiency). A problem with the latter does not preclude the former, but it does limit its potential in practice. Similarly, no one would condemn the field of mathematics as flawed simply because human minds are fallible and sometimes produce results for $2 + 2$ that differ from 4.[22]

Thus we fully recognise that changes are needed at the operational and research levels to help translate the intrinsic value associated with feature selectivity into practice in a robust manner. These efforts should include empirical studies of both examiner performance *and* the relative rarity (selectivity) of features, as well as a variety of improvements on systemic and conceptual dimensions. In the next section, we present and discuss specific challenges that researchers and practitioners should address and changes that should be introduced to place forensic feature comparison on a more appropriate basis.

Implications for future directions in research and practice

From what we have outlined in the previous sections, a number of discussion points arise regarding future directions for research and practice in forensic feature comparison. In this section, we address several of these discussion points. The order in which we present them does not reflect their importance. Furthermore, we do not claim that they represent an exhaustive list of considerations, nor that they are immediately or easily achievable. Our aim is to raise issues that we believe are critical to improving the trustworthiness of the discipline and, in the long term, to promoting a coherent understanding of it among researchers and recipients of expert evidence.

Overcoming the exclusive focus on examiner diagnosticism

From what we have explained in 'Points of disagreement', it should be clear that research programmes that focus *solely* on black-box testing of *examiners* and thereby characterising the general diagnosticity of expert utterances cannot get the discipline of forensic feature comparison and, more broadly, pattern-based examination off the ground at a fundamental level (Champod, 2014). As a reminder, in these research designs, examiners are given known ground-truth comparison pairs and then asked to give their opinion on whether or not the compared items are from the same source (source attribution

22. This analogy is inspired by Lindley's argument for normative over descriptive approaches to dealing with uncertainty (Lindley, 2017).

conclusions/identifications), or some ad hoc conclusion scale. It is disturbing to see how much effort is put into research using this design in a variety of disciplines, such as comparative handwriting examination (Hicklin et al., 2022), friction ridge skin mark (i.e., fingermark) examination (Growns and Kukucka, 2021; Ulery et al., 2011), forensic face comparison by so-called 'super-recognisers' (Hahn et al. 2022), tyre track examination (Richetelli et al., 2024) and, of interest to our discussion here, feature comparison as applied to firearms examination (Monson et al., 2022, 2023a, 2023b; Neuman et al., 2022), given how little—often nothing—such research contributes to fundamental understanding of the probative value of features in the first place.

A common characteristic of the category of studies mentioned above is that their design does not treat human examiners in a more sophisticated manner than other studies treat dogs to assess their olfactory detection ability (e.g., Guest et al., 2021; Marchal et al., 2016). The level of generality focuses primarily on (here: human) response data given a particular input stimulus, under varying experimental conditions. The resulting data are amenable, if at all, to mere summary statistics. As mentioned earlier in this article, exactly how such response data should be properly summarised is a highly contested issue that has generated so much controversy that external observers are left without clear guidance. As a result, in actual legal proceedings, some advocates tend to expose precisely this impasse because it may serve their case. However, we do not believe that this is the crux of the problem since similar objections can be raised on any issue where there is disagreement.

The deeper problem, as we see it, is that examiner diagnosticism leads to superficial and stereotypical objections that do not help recipients of expert information to assess the potential merits and limitations of feature comparison evidence *at a case-based level*. Furthermore, as long as there are researchers who use the examiner diagnosticism study scheme and elicit examiner responses about ground truth states (i.e., source attribution conclusions/identifications), they contribute to perpetuating the idea that it is appropriate for examiners to use a reporting format in which examiners directly opine on source propositions (Biedermann, 2022), despite the unscientific nature of such conclusions. As we will argue later in 'Abandoning the practice of source attribution conclusions (identification/individualisation)', we call for the abandonment of the source attribution conclusion format.

Quantifying the probative value of comparisons of marks and features

We cannot acquire, let alone claim, expertise in forensic feature comparison evidence until we address the difficult problem of quantifying the probative value of mark and feature comparisons. Even for identifications—which we do *not* endorse—this is an unavoidable preliminary step. However, the surface-level perspective of examiner diagnosticism, as pointed out in the previous section, is blind to this kind of deep understanding of the physical material being examined in the first place.

It is crucial to understand that scientific knowledge about the nature and selectivity of traces and marks is the basis for opinions. As a contrasting example, suppose that in forensic DNA we were to focus only on human response data as used in the examiner diagnosticism perspective, i.e., we were to ask the expert's opinion as to whether or not two examined biological traces, after comparing their DNA profiles, came from the same source. That would be nonsensical—no one would do that.[23] Forensic DNA experts are required to *quantify* the selectivity of the DNA features (e.g., STR markers) being compared. They must not opine on source-level propositions. But why, then, should source attribution opinions be acceptable in feature comparison disciplines? In our view, this is incoherent.

We sense that one answer to this may be that it is not possible to articulate the details of the examination process in feature comparison. For this reason, the PCAST report tells us, it is necessary to view the method as 'a "black box" in the examiner's head' (PCAST, 2016: 5), which brings us back to the

23. Except, maybe, for exclusions. But again, this kind of cases may not make it to trial.

superficial examiner diagnosticism perspective. We disagree. It is simply not true that the process of feature comparison is *completely* incomprehensible. For example, we know that there are different kinds of features: there are class characteristics (i.e., intentional design features) and there are other features that are subsequently acquired by different mechanisms (e.g., as a bullet travels through a barrel). We also have conceptual knowledge, i.e., a logical reasoning framework for evaluating class and acquired features together. This framework was described a quarter of a century ago in the context of shoemark evidence (Evett et al., 1998), and its logic also applies in the context of firearm evidence (Biedermann and Taroni, 2006).

It is true, however, that the challenge is to quantify the key components of this logical evaluation framework. But again, we know more than nothing. For example, we know that firearms vary widely in their class characteristics, and thus these have intrinsic probative value: as mentioned under 'Placing our critique into perspective', they allow one to reduce the population of potential sources (Champod and Biedermann, 2023). The next question is how much of a reduction, and how to justify a claimed reduction. These are difficult questions, but they should be addressed by research because—as noted above—they are the basis of an expert's opinion. In fact, we go further and suggest that a *justified* reduction factor (or, an expression of probative value), not just a judgement without an argument, should be considered a minimum requirement for an expert's opinion. In other words, *if* experts cannot provide and justify *at least* a qualitative (i.e., order of magnitude) reduction factor, they have no basis for giving an opinion. As a consequence, the features for which the selectivity cannot be justifiably quantified would have to be considered *uninterpreted* (Biedermann and Kotsoglou, 2022), and therefore not used to inform an examiner's opinion. We would certainly not recommend that, if an expert cannot justify at least a reduction factor, to simply ignore this fact and retreat to the perspective of examiner diagnosticism, which reduces the discussion to some sort of generic error rate. This would be tantamount to giving a pass to an examiner who has no thorough basis for his opinion.

In this regard, it should be noted that the value of the class characteristics has standing, where it can be asserted, regardless of whether one can assess the value of the acquired features. Stated otherwise, limitations in the inability to evaluate acquired features, e.g., striation marks, do not mean or imply that nothing can be said at all. The value of a mark depends on the level of detail (class vs. other types of characteristics) that is present and that the examiner is able to assess. Similarly, in fingermark examination, if all that can be retrieved from a fingermark is the general pattern, but no minutiae, it still makes sense to assess the reduction factor associated with the observed general pattern.[24]

The bottom line is that eliciting (quantifying) the probative value of forensic feature comparisons is challenging and requires a case-by-case approach. There is no predefined value for any feature comparison finding, and thus no simple recipe. However, forensic scientists are not left without hope: in an era where algorithmic approaches are ubiquitous, forensic scientists are better placed than ever to conduct research on feature selectivity (Swofford and Champod, 2021). At the same time, such research would open new avenues to mitigate the drawbacks of the current witness-centric perspective towards a more process-based perspective (Cheng and Nunn, 2019). Parts of the disciplines of fingermark examination (e.g., Neumann et al., 2012; Swofford et al., 2018), forensic voice comparison (e.g., Morrison et al., 2021; Ramos et al., 2021), and (firearms) feature comparison (e.g., Basu et al., 2022) demonstrate this. Unfortunately, the currently most prevalent perspective, which focuses only on the examiners' direct source attribution conclusions, turns a blind eye to these underlying complex interpretive challenges. Worse, keeping the focus *only* on general expert performance inevitably suggests to the field that addressing the deeper challenge of feature selectivity is not necessary or required.

24. This has led to the notion of non-identifiable fingermarks (Stoney et al., 2020).

Performing evidence interpretation

Our discussion so far might suggest that all that is needed to make the practice of feature comparisons more scientific is to quantify feature selectivity. But this would be somewhat shortsighted. To be trustworthy and scientifically robust, the field needs to embrace the full range of topics in forensic evidence interpretation (Aitken et al., 2010, 2020; Evett, 2015; Robertson et al., 2016; Willis et al., 2015). These include considerations such as asking the relevant questions (Stoney 1984), managing task-relevant information (National Commission on Forensic Science Human Factors Subcommittee, 2015), establishing the most relevant databases (Champod et al., 2004), and evaluating findings related to multiple levels of observation together, often referred to as 'combining evidence' (e.g., Juchli et al., 2012; Schum, 1994). Ultimately, scrutinising forensic feature comparisons at all of these levels represents a much broader, more comprehensive, and thus qualitatively better challenge than what the case-*un*specific focus on aggregate performance measures (e.g., error rates) associated with black-box testing can hope to achieve.

This is not to say that the PCAST report was unwise in advocating black-box testing for feature comparison disciplines, but it created a dilemma. Black-box testing was a reasonable proposal for *immediate* action to push the field to get at least some idea of examiner performance in general. However, the proposal is not well suited to a long-term perspective. Indeed, it should not be taken to mean that from now on it would be sufficient to conduct *only* black-box testing without making efforts to quantify feature selectivity. Even the PCAST report recognises the importance of being able to assign probative value to features: '(…) the expert should report the probative value (…) based on the specific features observed in the case' and '[t]he frequency with which a particular pattern or set of features will be observed in different samples (…) is an essential element in drawing conclusions' (PCAST, 2016: 6). This is precisely what we advocate throughout this Commentary, particularly under 'Quantifying the probative value of comparisons of marks and features'.

Abandoning the practice of source attribution conclusions (identification/individualisation)

As a necessary step towards achieving a more scientific status, forensic feature comparison experts should abandon the practice of providing source attribution conclusions (identification/individualisation), in favour of scientifically defensible assessments of the value of evidence. This transition to the post-identification era is long overdue. Over the past three decades, various commentators have used different but complementary arguments to point out the unscientific nature of categorical source attribution conclusions (e.g., Biedermann et al., 2008; Saks and Koehler, 2005; Stoney, 1991, 2012). For example, the claim inherent in source attribution conclusions that all other potential sources are excluded is an overstatement that goes beyond what the available findings can demonstrate. There is simply no *empirical* demonstration of the exclusion of all other potential sources, except perhaps in closed-set situations where an exhaustive examination of all candidate sources can be performed and demonstrated. But even in that situation, the conclusion of identification would require the strong assumption of an error-free examination process.

If professional forensic organisations were genuinely committed to science, they would adopt this proposed change in reporting practice *today*, or at least prepare a roadmap for implementing a change. This would contribute to advancing the scientification of forensic practice (Biedermann, 2022; Koehler et al., 2023).

Our call to abandon source attribution conclusions also includes weaker formats that use the term identification *without* going so far as to assert the exclusion of all other sources. An example is the Uniform Language for Testimony and Reports for the Forensic Firearms/Toolmarks Discipline (Pattern Examination) issued by the U.S. Department of Justice, which allows examiners to make a 'source

identification,' defined as 'an examiner's conclusion that two toolmarks originated from the same source' (U.S. Department of Justice, 2023: 2). At the same time, the document requires that 'an examiner shall not (…) assert that two toolmarks originated from the same source to the exclusion of all other sources' (U.S. Department of Justice, 2023: 3). Not only is this contradictory, but it also runs the danger of misleading recipients of expert conclusions—mesmerised by the categorical flair of the term 'identification'—into believing that the conclusion actually excludes all other sources.

Using proper terminology

It is difficult to see how one can properly think about and understand the possibilities and limitations of forensic feature comparison if vague and confusing terminology continues to be used. Specifically, we believe that there are at least two terms, 'match(ing)' and 'decision (making)', that researchers and practitioners should avoid.

First, consider the term 'match'. This term is inappropriate in a number of ways. For example, some use the term to describe both the findings (observations) of a feature comparison and the ground truth (i.e., whether or not two compared items come from the same source).[25] This is confusing to the recipients of expert information, who may not know exactly what is meant (observation or ground truth).[26] Furthermore, there is the potential for misunderstanding or misconstruing the summary of an observation in terms of a match as a conclusion that the compared items come from the same source, even though this is logically incorrect. After all, the term 'match' is similar to, or may even suggest equivalence to, the term identity. However, two compared objects cannot perfectly match, i.e., be identical, because—by definition—an object can only be identical to itself. In this sense, two objects, even if they come from the same source, can *never* match (i.e., be identical). At best, two objects may be indistinguishable from each other *at the chosen level of observation*, and this in itself tells us nothing about whether they come from the same source or not. There are many objects that are indistinguishable (or are found to match) at the chosen level of observation, but come from different sources.

The term 'decision' is the second term that is a source of widespread confusion. It is often used in conjunction with the term 'match' to refer to the conclusion reached by an examiner and/or the process that led to the examiner's conclusion. For example, Smith and Neal (2021) refer to 'forensic procedures that involve making "match" decisions between a crime-scene sample and a sample from the suspect' (2021: 1). Similarly, over the past decade, the term has largely seduced researchers in a variety of disciplines. Examples include the following article titles [emphasis added]: 'How to make better forensic *decisions*' (Albright, 2022), 'Accuracy of comparison *decisions* by forensic firearms examiners' (Monson et al., 2023a), 'Accuracy and reliability of forensic latent fingerprint *decisions*' (Ulery et al., 2011). The use of the term 'decision' in these publications is inappropriate, firstly because it conveys a kind of sophistication or exceptionalism that, as we have argued throughout this Commentary, source attributions conclusions lack. Put another way, the term 'decision' mislabels and misrepresents the nature of forensic source attributions by its tendency to distract from and obscure their unscientific character. Second, examining forensic source attributions through the lens of decision theory (Biedermann et al., 2008) actually leads to the conclusion that, in practice, forensic examiners cannot legitimately (claim to) make identification conclusions (Cole and Biedermann, 2020; Kotsoglou and Biedermann, 2022; Stoney, 2012). The bottom line is that while it may seem that, *descriptively*, the term 'decision' summarises what examiners think they are doing, analytically and conceptually this is not the case.

25. See Growns and Kukucka (2021), Scurich and John (2023) and Smith and Neal (2021) for examples and Biedermann (2022) for a critical discussion.
26. The lack of definitional clarity can also lead to confusion in formulaic developments.

Giving up forensic science exceptionalism

Large parts of forensic science exhibit a longstanding tendency to focus on self-centred disciplinary discourses about the field's legitimacy and definition, with self-referential arguments about training and experience rather than classical research principles (Fabricant, 2022; Koehler, 2017; Mnookin et al., 2012; Risinger and Saks, 2003; Saks and Faigman, 2008).[27] Such discourses tend to lead to spurious *exceptionalism* (e.g., Biedermann and Kotsoglou, 2020; Edmond and Martire, 2018; Faigman, 2008; Saks, 2007), which we believe is one of the reasons why parts of forensic science suffer from a lack of scientification (Koehler et al., 2023). In some disciplines, particularly outside forensic genetics,[28] forensic scientists have claimed and continue to claim that they are entitled to conclusions of absolute certainty because there is something seemingly special about their 'science', without actually being able to articulate what that specialness is. This attitude is accompanied by a certain disrespect for other branches of science, such as statistics[29] and, as Faigman et al. (2022) point out, experimental and research design. The implication or belief is that forensic science can do without these traditional, hardwired branches of science.

This dismissive attitude is hardly surprising, given that a conceptual analysis of the core problems of forensic science reveals that one of the discipline's most cherished products, source attribution conclusions, is scientifically unsound. Already more than 30 years ago, Stoney (1991) noted that there is no magic in forensic source attribution conclusions. Instead, by going beyond what is warranted by the available evidence, such conclusions require 'a leap of faith' (Stoney, 1991: 198), which makes them unscientific. And yet, recent reviews have found that identification conclusions are still firmly in place (Swofford et al., 2021).

To break this vicious circle of problematic practices, forensic science should abandon exceptionalism and isolationism. It must be open to all branches of science that can help serve the interests of justice. In addition, forensic science should not only acknowledge its limitations and the need to collaborate with other scientific disciplines, but also commit to doing so (Koehler et al., 2023). Otherwise, forensic science will remain its own stumbling block in the feature comparison debate. In summary, therefore, we advocate a broadening of interests and perspectives, i.e., a critical assessment of what exactly *each* science can and cannot legitimately contribute to the purpose of legal evidence and proof processes.

Discussion and conclusions

Current research and practice in the field of forensic feature comparison as applied to firearms examination is widely regarded as problematic. Of primary concern are identification and individualisation conclusions, i.e., testimonial claims that the pool of potential sources of a forensic trace can be reduced to a single source (to the exclusion of all others). These conclusions are paradoxical because they persist despite the fact that never before in the history of forensic science have we had a better understanding of the limitations, irrationality and unscientific nature of such claims. In both research and professional practice, this persistent irrationality is a source of the disrespect and distrust that members of the judiciary and the public have for forensic science (Fabricant, 2022). It is a failure of forensic science and a reminder that the post-identification era is long overdue.

27. This attitude is well illustrated by the responses of the forensic science community and stakeholders to the PCAST report and the subsequent 2017 addendum to that report (PCAST 2017).
28. Incidentally, and in contrast to other forensic disciplines, forensic genetics is perhaps the field best positioned to legitimately claim exceptionalism (Murphy, 2009), but even in this field, conclusions implying certainty are the exception, *not* the rule.
29. Professor James Curran, for example, has asked whether 'forensic science [is] the last bastion of resistance against statistics' (Curran, 2013).

In light of this, several changes to current research and practice are needed to help the field improve its scientific foundation and trustworthiness. These changes should primarily include a move towards more defensible assessment and reporting procedures that avoid categorical conclusions about source attribution (Koehler et al., 2023). Successful implementation of these changes will require widespread commitment.

As forensic feature comparison as applied to firearms examination is currently under serious challenge from various non-forensic circles as well as from the courts, this challenge is most welcome. It has gained momentum and strongly suggests that it is here to stay (Garrett et al., 2023), reinforcing our point that change is now not only long overdue, but inevitable.

We broadly agree with the mainstream criticism of forensic feature comparison: the traditional and currently most common way of conducting and reporting forensic feature comparisons, as applied in the context of firearms examination, is unacceptable. However, we disagree in part with the reasons for this assessment and the way forward. The critics point to the lack of evidence of examiner proficiency in general and condemn poor performance in those cases where there is little evidence on the matter. This is a valid criticism and a relevant line of attack in actual admissibility proceedings. However, as we have pointed out, we consider the exclusive focus on state-of-the-art examiner performance (i.e., examiner diagnosticity) to be a one-sided and superficial criticism. While the trustworthiness and capacity of the field does depend on the proficiency of examiners, the field is not reducible to it. Examiner diagnosticism ignores and offers no way to strengthen the scientific understanding of the intrinsic value of marks and traces (i.e., feature selectivity). As noted under 'Overcoming the exclusive focus on examiner diagnosticism', it is no exaggeration to say that the exclusive focus on examiner diagnosticity is conceptually no better than trying to make sense of the wagging tails of sniffer dogs.

In the long term, the current and exclusive emphasis on the examiner diagnosticism perspective, based primarily on black-box studies such as those advocated in the PCAST report, is a dead end for two reasons. First, in the future, scientists should no longer make categorical source attributions conclusions (identification/individualisation) (Biedermann, 2022; Kaye, 2023; Koehler et al., 2023; Morrison, 2022), thus removing an essential object of purely descriptive research concerned with surveying examiners' opinions about test comparison pairs and debates about how such data should be summarised. Second, as even the PCAST report acknowledges, an examiner's conclusion should be based on a deep understanding of the process of feature generation, the nature of features, and their informative value. Therefore, moving away from direct identification conclusions will force scientists to think about how to quantify the value of their observations (feature comparisons) and measurements, thus bringing the feature selectivity perspective to the forefront. Empirical testing will continue to be a relevant part of the future, but the subject of these studies should not be the mere opinions of forensic examiners, but measurement and evaluation procedures that produce value of evidence assessments based on measurable features (Aitken et al., 2020; Meuwly et al., 2017; Morrison et al., 2021; Ramos and Gonzalez-Rodriguez, 2013; Ramos et al., 2021).

Acknowledgments

The authors are grateful to two anonymous referees for their comments which helped to improve the manuscript. Constructive comments were also provided by Hillel Bavli, David Caudill, Christopher Lau and Alexander Walker III during the Evidence Summer Workshop (ESW2024), Vanderbilt Law School, Nashville, TN (15–17 May 2024).

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The first author gratefully acknowledges the support of the Swiss Benevolent Society of New York. The authors thank the Consortium of Swiss Academic Libraries for supporting the open access publication of this article.

ORCID iD


Alex Biedermann https://orcid.org/0000-0002-0271-5152

References

Aitken CGG, Roberts P and Jackson G (2010) Fundamentals of Probability and Statistical Evidence in Criminal Proceedings (Practitioner Guide No. 1), Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses, Royal Statistical Society's Working Group on Statistics and the Law.

Aitken CGG, Taroni F and Bozza S (2020) *Statistics and the Evaluation of Evidence for Forensic Scientists*, 3rd ed. Chichester: John Wiley & Sons.

Albright TD (2022) How to make better forensic decisions. *Proceedings of the National Academy of Sciences* 119: e2206567119.

Basu N, Bolton-King RS and Morrison GS (2022) Forensic comparison of fired cartridge cases: Feature-extraction methods for feature-based calculation of likelihood ratios. *Forensic Science International: Synergy* 5: 100272.

Biedermann A (2022) The strange persistence of (source) 'identification' claims in forensic literature through descriptivism, diagnosticism and machinism. *Forensic Science International: Synergy* 4: 100222.

Biedermann A, Bozza S and Taroni F (2008) Decision theoretic properties of forensic identification: Underlying logic and argumentative implications. *Forensic Science International* 177: 120–132.

Biedermann A and Kotsoglou K (2020) Digital evidence exceptionalism? A review and discussion of conceptual hurdles in digital evidence transformation. *Forensic Science International: Synergy* 2: 262–274.

Biedermann A and Kotsoglou K (2022) (Un-)interpretability in expert evidence: An inquiry into the frontiers of evidential assessment. *Quaestio Facti (Revista Internacional Sobre Razonamiento Probatorio Quaestio Facti. International Journal on Evidential Legal Reasoning* 3: 481–515.

Biedermann A and Kotsoglou KN (2021) Forensic science and the principle of excluded middle: 'inconclusive' decisions and the structure of error rate studies. *Forensic Science International: Synergy* 3: 100147.

Biedermann A and Taroni F (2006) A probabilistic approach to the joint evaluation of firearm evidence and gunshot residues. *Forensic Science International* 163: 18–33.

Biedermann A and Vuille J (2018) The decisional nature of probability and plausibility assessments in juridical evidence and proof. *International Commentary on Evidence* 16: 1–30.

Burnell R, Schellaert W, Burden J, et al. (2023) Rethink reporting of evaluation results in AI. *Science* 380: 136–138.

Champod C (2014) Research focused mainly on bias will paralyse forensic science. *Science & Justice* 54: 107–109.

Champod C and Biedermann A (2023) Identification/individualisation, overview and meaning of ID. In: Houck MM (eds) *Encyclopedia of Forensic Sciences* (3rd ed). Oxford: Elsevier, 53–62.

Champod C, Evett IW and Jackson G (2004) Establishing the most appropriate databases for addressing source level propositions. *Science & Justice* 44: 153–164.

Champod C, Lennard C, Margot P, et al. (2016) *Fingerprints and Other Ridge Skin Impressions*, 2nd ed. Boca Raton: CRC Press.

Cheng EK and Nunn GA (2019) Beyond the witness: Bringing a process perspective to modern evidence law. *Texas Law Review* 97: 1077–1124.

Cole SA and Biedermann A (2020) How can a forensic result be a 'decision'? A critical analysis of ongoing reforms of forensic reporting formats for federal examiners. *Houston Law Review* 57: 551–592.

Committee for the Advancement of the Science of Firearm & Toolmark Identification (2011) Theory of identification as it relates to toolmarks: Revised. *AFTE Journal* 43: 287.

Curran JM (2013) Editorial: Is forensic science the last bastion of resistance against statistics? *Science & Justice* 53: 251–252.

Edmond G and Martire K (2018) Antipodean forensics: A comment on ANZFSS's response to PCAST. *Australian Journal of Forensic Sciences* 50: 140–151.

Evett IW (2015) The logical foundations of forensic science: Towards reliable knowledge. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 370: 1–10.

Evett IW, Lambert JA and Buckleton JS (1998) A Bayesian approach to interpreting footwear marks in forensic casework. *Science & Justice* 38: 241–247.

Fabricant CM (2022) *Junk Science and the American Criminal Justice System*. Brooklyn: Akashic Books.

Faigman DL (2008) Anecdotal forensics, phrenology, and other abject lessons from the history of science. *Hastings Law Journal* 59: 979–1000.

Faigman DL, Monahan J and Slobogin C (2014) Group to individual (G2i) inference in scientific testimony. *The University of Chicago Law Review* 81: 417–480.

Faigman DL, Scurich N and Albright TD (2022) The field of firearms forensics is flawed. *Scientific American*, May 25.

Garrett BL, Scurich N, Tucker E, et al. (2023) Judging firearms evidence and the rule 702 amendments. *Judicature* 107: 41–49.

Gastwirth JL (1987) The statistical precision of medical screening procedures: Application to polygraph and AIDS antibodies test data. *Statistical Science* 2: 213–222.

Growns B and Kukucka J (2021) The prevalence effect in fingerprint identification: Match and non-match base-rates impact misses and false alarms. *Applied Cognitive Psychology* 35: 751–760.

Guest C, Harris R, Sfanos KS, et al. (2021) Feasibility of integrating canine olfaction with chemical and microbial profiling of urine to detect lethal prostate cancer. *PLoS ONE* 16: 1–23.

Guyll M, Madon S, Yang Y, et al. (2023) Validity of forensic cartridge-case comparisons. *Proceedings of the National Academy of Sciences* 120: e2210428120.

Hahn CA, Tang LL, Yates AN, et al. (2022) Forensic facial examiners versus super-recognizers: Evaluating behavior beyond accuracy. *Applied Cognitive Psychology* 36(6): 1209–1218.

Hicklin RA, Eisenhart L, Richetelli N, et al. (2022) Accuracy and reliability of forensic handwriting comparisons. *Proceedings of the National Academy of Sciences* 119: e2119944119.

Imwinkelried EJ (2020) The admissibility of scientific evidence: Exploring the significance of the distinction between foundational validity and validity as applied. *Syracuse Law Review* 70: 817–849.

Juchli P, Biedermann A and Taroni F (2012) Graphical probabilistic analysis of the combination of items of evidence. *Law, Probability and Risk* 11: 51–84.

Kaye DH (1987) The validity of tests: Caveant omnes. *Jurimetrics Journal* 27: 349–361.

Kaye DH (2023) Maryland Supreme Court resists 'unqualified' firearms-toolmark testimony. Available at https://for-sci-law.blogspot.com/2023/06/maryland-supreme-court-resists.html.

Kirk PL (1963) The ontogeny of criminalistics. *Journal of Criminal Law, Criminology and Police Science* 54: 235–238.

Koehler JJ (2008) Fingerprint error rates and proficiency tests: What they are and why they matter. *Hastings Law Journal* 59: 1077–1100.

Koehler JJ (2017) Forensics or fauxrensics? Ascertaining accuracy in the forensic sciences. *Arizona State Law Journal* 49: 1369–1416.

Koehler JJ, Mnookin JL and Saks MJ (2023) The scientific reinvention of forensic science. *Proceedings of the National Academy of Sciences* 120: 1–24.

Kotsoglou KN and Biedermann A (2022) Inroads into the ultimate issue rule? Structural elements of communication between experts and fact finders. *The Journal of Criminal Law* 86: 223–240.

Lindley DV (2017) Foreword. In: de Finetti B (eds) *Theory of Probability, A Critical Introductory Treatment*, Reprint ed. Chichester: John Wiley & Sons, ix–xi.

Marchal S, Bregeras O, Puaux D, et al. (2016) Rigorous training of dogs leads to high accuracy in human scent matching-to-sample performance. *PLoS ONE* 1: e0146963.

Meuwly D, Ramos D and Haraksim R (2017) A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic Science International* 276: 142–153.

Mnookin JL, Cole SA, Dror IE, et al. (2012) The need for a research culture in the forensic sciences. *UCLA Law Review* 58: 725–779.

Monson KL, Smith ED and Bajic SJ (2022) Planning, design and logistics of a decision analysis study: The FBI/Ames study involving forensic firearms examiners. *Forensic Science International: Synergy* 4: 100221.

Monson KL, Smith ED and Peters EM (2023a) Accuracy of comparison decisions by forensic firearms examiners. *Journal of Forensic Sciences* 68: 86–100.

Monson KL, Smith ED and Peters EM (2023b) Repeatability and reproducibility of comparison decisions by firearms examiners. *Journal of Forensic Sciences* 68: 1721–1740.

Morrison GS (2022) A plague on both your houses: The debate about how to deal with 'inconclusive' conclusions when calculating error rates. *Law, Probability and Risk* 21: 127–129.

Morrison GS, Enzinger E, Huges V, et al. (2021) Consensus on validation of forensic voice comparison. *Science & Justice* 61: 299–309.

Morrison GS, Kaye DH, Balding DJ, et al. (2017) A comment on the PCAST report: Skip the 'match'/'non-match' stage. *Forensic Science International* 272: e7–e9.

Murphy E (2009) What 'strengthening forensic science' today means for tomorrow: DNA exceptionalism and the 2009 NAS report. *Law, Probability and Risk* 9: 7–24.

National Commission on Forensic Science Human Factors Subcommittee (2015) Views of the commission: Ensuring that forensic analysis is based upon task-relevant information.

Neuman M, Hundl C, Grimaldi A, et al. (2022) Blind testing in firearms: Preliminary results from a blind quality control program. *Journal of Forensic Sciences* 67: 964–974.

Neumann C, Evett IW and Skerrett J (2012) Quantifying the weight of evidence from a fingerprint comparison: A new paradigm. *Journal of the Royal Statistical Society, Series A* 175: 371–416.

PCAST (2016) *President's Council of Advisors on Science and Technology*, *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. Washington, D.C.: Executive Office of the President.

PCAST (2017) President's Council of Advisors on Science and Technology, An Addendum to the PCAST Report on Forensic Science in Criminal Courts. Washington, D.C. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast forensics addendum finalv2.pdf.

Ramos D and Gonzalez-Rodriguez J (2013) Reliable support: Measuring calibration of likelihood ratios. *Forensic Science International* 230: 156–169.

Ramos D, Meuwly D, Haraksim R, et al. (2021) Validation of forensic automatic likelihood ratio methods. In: Banks DL, Kafadar K, Kaye DH and Tackett M (eds) *Handbook of Forensic Statistics*. Boca Raton: CRC Press, 143–163.

Richetelli N, LeMay J, Dunagan KM, et al. (2024) Accuracy and reproducibility of forensic tire examination decisions. *Forensic Science International* 358: 112009.

Risinger DM and Saks M (2003) Rationality, research and leviathan: Law enforcement-sponsored research and the criminal process. *Michigan State DCL Law Review* 4: 1023–1050.

Robertson B, Vignaux GA and Berger CEH (2016) *Interpreting Evidence. Evaluating Forensic Science in the Courtroom*, 2nd ed. Chichester: John Wiley & Sons.

Rosenblum M, Chin ET and Ogburn EL, (2024) Misuse of statistical method results in highly biased interpretation of forensic evidence in Guyll et al. (2023). *Law, Probability and Risk* 23, (in press).

Saks MJ (2007) Remediating forensic science. *Jurimetrics* 48: 119–124.

Saks MJ and Faigman DL (2008) Failed forensics: How forensic science lost its way and how it might yet find it. *Annual Review of Law and Social Science* 4: 149–171.

Saks MJ and Koehler JJ (2005) The coming paradigm shift in forensic identification science. *Science* 309: 892–895.

Schum DA (1994) *Evidential Foundations of Probabilistic Reasoning*. New York: John Wiley & Sons, Inc.

Schwartz A (2005) A systematic challenge of the reliability and admissability of firearms and toolmarks identification. *The Columbia Science and Technology Law Review* 6: 1–42.

Scurich N and John RS (2023) Three-way ROCs for forensic decision making. *Statistics and Public Policy* 10: 1–10.

Smith AM and Neal TMS (2021) The distinction between discriminability and reliability in forensic science. *Science & Justice* 61: 319–331.

Stoney DA (1984) Evaluation of associative evidence: Choosing the relevant question. *Journal of the Forensic Science Society* 24: 473–482.

Stoney DA (1991) What made US ever think we could individualize using statistics? *Journal of the Forensic Science Society* 31: 197–199.

Stoney DA (2012) Discussion on the paper by Neumann, Evett and Skerrett. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 175: 399–400.

Stoney DA, De Donno M, Champod C, et al. (2020) Occurrence and associative value of non-identifiable fingermarks. *Forensic Science International* 309: 110219.

Swofford H and Champod C (2021) Implementation of algorithms in pattern & impression evidence: A responsible and practical roadmap. *Forensic Science International: Synergy* 3: 100142.

Swofford HJ, Cole SA and 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.

Swofford HJ, Koertner AJ, Zemp F, et al. (2018) A method for the statistical interpretation of friction ridge skin impression evidence: Method development and validation. *Forensic Science International* 287: 113–126.

Swofford HJ, Lund S, Iyer H, et al. (2024) Inconclusive decisions and error rates in forensic science. *Forensic Science International: Synergy* 8: 100472.

Taroni F, Biedermann A, Bozza S, et al. (2014) *Bayesian Networks for Probabilistic Inference and Decision Analysis in Forensic Science*, 2nd ed. Statistics in Practice. Chichester: John Wiley & Sons.

Taroni F, Biedermann A, Garbolino P, et al. (2004) A general approach to Bayesian networks for the interpretation of evidence. *Forensic Science International* 139: 5–16.

Thompson WC, Taroni F and Aitken CGG (2003) How the probability of a false positive affects the value of DNA evidence. *Journal of Forensic Sciences* 48: 47–54.

Ulery B, Hicklin A, Buscaglia J, et al. (2011) Accuracy and reliability of forensic latent fingerprint decisions. *Proceedings of the National Academy of Science of the United States of America* 108: 7733–7738.

U.S. Department of Justice (2023) Uniform Language for Testimony and Reports for the Forensic Firearms/Tool- marks Discipline (Pattern Examination), vers. 5.18.23. Available online at: https://www.justice.gov/d9/2023- 05/firearms pattern examination ultr 5.18.23.pdf.

Willis S, McKenna L, McDermott S, et al. (2015) ENFSI Guideline for Evaluative Reporting in Forensic Science. *Strengthening the Evaluation of Forensic Results Across Europe (STEOFRAE). Dublin*.