POSTPRINT – Published as:
Biedermann A. 2026, 'Inconclusives': Divide and conquer?—Commentary on "A Variance Decomposition Approach to Inconclusives in Forensic Black Box Studies" by A. Luby and J. Kadane (2026-04-30), Law, Probability and Risk, 2026, 25, 1–7, https://doi.org/10.1093/lpr/mgag002

INVITED COMMENTARY

'Inconclusives': Divide and Conquer? – Commentary on "A Variance Decomposition Approach to Inconclusives in Forensic Black Box Studies" by A. Luby and J. Kadane

*Alex Biedermann**

University of Lausanne, Faculty of Law, Criminal Justice and Public Administration School of Criminal Justice, 1015 Lausanne–Dorigny (Switzerland)

1. Introduction

As the assessment of performance at the level of both individual forensic examiners and entire forensic disciplines is a constant concern, approaches to analysing, summarising and interpreting data from empirical studies (i.e. 'black box' or 'validation' studies) are the subject of ongoing debate. Much of the controversy relates to the idea that error rates should characterise the performance of forensic examiners conducting comparison tasks and expressing their conclusions as 'identification', 'exclusion' or 'inconclusive'. However, it is unclear how such error rates should be calculated, not to mention how they should be understood and used by the judiciary. Additionally, there is an intense debate about whether the 'inconclusive' conclusion category should be considered an error, since this has an impact on the resulting error rates.

The paper by Luby and Kadane (2025) is a valuable and welcome contribution to this ongoing discussion. Section 2 of this commentary briefly outlines the novel aspects of Luby and Kadane's (2025) paper, while Section 3 discusses specific points raised by the authors. Section 4 concludes by presenting broader considerations, critical aspects and remaining problems in the field.

2. Progress in dealing with 'inconclusives' in forensic black box study data

There are three aspects that distinguish Luby and Kadane's (2025) paper from other recent discussions on how to handle 'inconclusives' in data from forensic black box studies.

Firstly, the authors' analysis is original in that they do not immediately focus on the notion of error rate and how 'inconclusives' should, if at all, be incorporated into this notion. Instead, they encourage readers to take a step back and look 'behind the curtain'. They convincingly argue that insight can be gained by investigating what they term the *pattern* of 'inconclusives' in black box studies. This can reveal differences between studies that may not be apparent when focusing solely on error rates as a single summary of study data. For example, it can be informative to know the proportion of 'inconclusive' responses that each examined item (or pair of items) has received, as well as the proportion of 'inconclusive' conclusions that each examiner has reached. The former, the proportion of 'inconclusive' conclusions reached on an item, informs us about its difficulty or *assertability*. The latter, the proportion of an examiner's conclusions that are 'inconclusives' can, as noted previously in Biedermann and Kotsoglou (2021), be considered a descriptive measure of an examiner's *assertiveness*. Luby and Kadane (2025) thus emphasise, in addition to considering summary statistics at the level of studies as a whole, that there is also value in isolating particular components of a study, such as the 'inconclusive' response category. In terms of methodology, this amounts to a 'divide and conquer' approach whereby studies are considered in terms of their components before any

* Email: alex.biedermann@unil.ch, ORCID: https://orcid.org/0000-0002-0271-5152

overall summary statistics are examined. Swofford et al. (2024) recently made a similar point, cautioning against aggregating validation study data into summary statistics due to the associated information loss and inadequacy for subsequent rational inference and decision-making.

Secondly, and as a corollary to the above, Luby and Kadane (2025) offer a new perspective at a time when previous discussions have seemingly reached an impasse. Their approach of considering study components separately first is more rigorous and disciplined than previous perspectives, which seem to assume that deriving a single overall statistic (often with few or even no explanations of the underlying assumptions) is the primary goal. Luby and Kadane (2025) validly point out that deriving such an overall statistic is an important and informative process in itself and not just a means to an end. Indirectly, they thus warn against the possible drawbacks of reductionism and oversimplification associated with the uniform treatment of 'inconclusives' to date. In other words, the authors' message is that 'inconclusive' conclusions are more than just counts. Instead, they offer additional dimensions by which black box studies can be characterised and compared, both within and across disciplines.

Thirdly, Luby and Kadane's (2025) methodological proposal, illustrated using two existing studies, demonstrates the added value that forensic statistics can bring to the discussion when validation study data are recorded and made available at the level of individual examiners. In other words, forensic statistics is not an optional nicety to consider once data are available; it is a necessary discipline from the outset, as Curran (2012) pointed out earlier.

3. Specific points raised in the analysis by Luby and Kadane (2025)

One of the most distinctive features of Luby and Kadane's (2025) approach to 'inconclusives' is that they do not treat this category of conclusion in a uniform manner as either correct, incorrect, or ignorable. This perspective is welcome insofar as 'inconclusive' indeed does not logically map onto the traditional notion of error which is defined as a discrepancy between an examiner's statement and the actual ground truth. Instead, an examiner's 'inconclusive' statement is better viewed as a void, i.e. the examiner *abstains* from providing one of the categorical expressions 'same source' or 'different source' that could be meaningfully compared with the ground truth and labelled as accurate or erroneous. Labelling an 'inconclusive' response as a potential error would require one to assume that the examiner *should* have provided a different answer. This would introduce an additional element to the problem in the form of a value judgment regarding the *desirable* response in the case at hand. However, as Arkes and Koehler (2021) have noted, this is essentially unresolvable.

While Luby and Kadane (2025) are careful not to label *individual* inconclusive conclusions as correct, incorrect or ignorable, they adopt a slightly different stance when characterising 'inconclusives' at the level of studies as a whole. Here, the authors state that they "weight some proportion of the inconclusives as attributable to examiners (and therefore potential errors)", thus leaning towards the idea that an 'inconclusive' could be an error. This raises the question of whether Luby and Kadane (2025) create a back door for the idea of an 'inconclusive' being a potential error. However, how could the notion of an erroneous 'inconclusive' emerge at the level of studies as a whole if the concept of an error in an individual 'inconclusive' report is undefined? This is a complicated question, and Luby and Kadane (2025) emphasise carefully that their summary statistic for studies as a whole is a "failure rate", *not* an "error rate". While this level of semantic and conceptual nuance is commendable, it cannot prevent new questions from arising, such as the difference between failure and error rates and how end users should deal with these concepts.

Besides the question of whether an 'inconclusive' can be erroneous, there is the more fundamental question of what an 'inconclusive' actually is. In relation to this question, it is worth noting that Luby and Kadane (2025) focus on "(…) the overall patterns of inconclusives in each particular study to weight some proportion of the inconclusives as attributable to examiners (and therefore potential errors) and some proportion as attributable to the items (and therefore attributable to the study designers)". The authors also state that they "(…) write with the premise that an inconclusive determination arises from the interaction between the examiner and the item, and hence reflects both".

The authors make two subtle points here that require careful consideration. Firstly, by writing "inconclusives (…) as *attributable* to the items" (*emphasis* added) the authors suggest that 'inconclusive' is in part a *property* of an item. At the same time, by stating "inconclusives as *attributable* to examiners" (*emphasis* added), they also consider 'inconclusive' to be a characteristic of an examiner. Secondly, by stating that "an inconclusive determination *arises from*" (*emphasis* added), the authors divert the reader's attention from the *nature* of an 'inconclusive' to the *factors* considered relevant for assessing the probability of an examiner reporting 'inconclusive'.

By making these assertions, the authors avoid addressing the fundamental question of what an 'inconclusive' is, at least in the initial parts of their paper. Instead, they describe the process and circumstances under which an 'inconclusive' arises. The problem with this way of conceptualising 'inconclusives' is that it tends to carry the term 'inconclusive' as a descriptor of a reporting category over to a description – *inconclusiveness* – of an examined item. This inevitably brings the author's account close to the modified but flawed error rate study design of Dror and Scurich's (2020), in which 'inclusive' is illogically treated as a third ground truth state (alongside 'same source' and 'different source').

Unfortunately, this creates ambiguity and confusion where clarity is required. Therefore, it is helpful to recall that, by definition, 'inconclusive' is a conclusion category used by an examiner: it designates an examiner's utterance. Nothing more, nothing less. 'Inconclusive' is *not* an item's property.

This does not deny that some items may be more challenging for an examiner than others. In other words, examiners may find it more or less difficult to reach a definite conclusion as to whether the items are from the same source or different sources. However, this should not lead us to believe or to suggest – as Dror and Scurich (2020) did – that an *item* is inconclusive. Items may pose difficulties for various reasons, primarily due to the quality and quantity of recognisable features. Therefore, at item level, we are concerned with *assertability* (or 'decideability'), not inconclusiveness. Various metrics currently exist for characterising the quality of items (i.e. their suitability for analysis), depending on the type of trace considered (e.g., fingermarks).

To further illustrate the above distinction, it is helpful to consider formal notation. In the traditional reporting scheme, denote by R the *event* of an examiner reporting one of the following mutually exclusive expressions 'identification', 'exclusion' or 'inconclusive'. Denote the possible ground truth states 'same source' and 'different source' by H. In addition, consider a variable D, representing a measure of item difficulty. Based on these definitions, we can consider the probability Pr of the event that an examiner will report 'inconclusive' (R), given the ground truth state (H), and that the pair of items being compared are of a given level of difficulty (D). More formally, this can be written as $\Pr(R \mid H, D)$.[1] Clearly, in this expression,

[1] Note that this is a simple way of formalising the thinking about 'inconclusives'. Luby and Kadane (2025) introduce a more elaborate hierarchical model, considering an examiner's tendency to be assertive as an additional conditioning variable.

only the variable *R* should represent the concept of 'inconclusive', as given by the definition of this variable. Mixing it with either *H* or *D* would amount to conditioning a variable (here, *R*) by itself, leading to circular reasoning.

The formal notation introduced above also shows that there is a difference between the *event* of a reported 'inconclusive' and our *probability* of observing such an event. Therefore, Luby and Kadane's (2025) statement that "an inconclusive determination arises from the interaction between the examiner and the item, and hence reflects both" could be restated in probabilistic terms as follows: "Our *probability* of observing a given examiner reporting 'inconclusive' *depends* on the examiner's assertiveness and the assertability (difficulty) of the examined item". This way of thinking about 'inconclusive' is important because it helps us avoid mischaracterising the notion of 'inconclusive': strictly speaking, it is *not* the 'inconclusive' (determination) that is a function of other variables. Instead, mathematically speaking, it is our *probability* of the uncertain event of observing the report of an 'inconclusive' that is a function of our knowledge (i.e., other variables) at the time we express uncertainty.

Interestingly, Luby and Kadane (2025) appear to subscribe to this perspective in later parts of their paper. In the section "Model-Based Variance Decomposition", they present a model in which the *probability* of an examiner reporting inconclusive for a given item is a function of two variables that they call "participant proficiency" and "item easiness". By using the expression "item easiness", the authors clearly refer to the difficulty or assertability of an item, thus avoiding any suggestion that 'inconclusive' is a property of an item.

4. Pending challenges in the assessment of the performance of forensic practitioners

While Luby and Kadane's (2025) paper improves our understanding of how to analyse black box study data and advances conceptual precision in the current debate over 'inconclusives', it is inevitably subject to fundamental limitations resulting from the structure of traditional error rate studies. However, it should be noted that these limitations affect all current approaches to such studies, and are not a fault of the authors' paper. To provide further context, it is worth mentioning some of the conceptual and practical problems that characterise the topic of 'inconclusives' more broadly. For readers unfamiliar with the controversy surrounding 'inconclusive', it is important to explain why these problems warrant attention and reiteration.

Start by considering the notion error (rate). What exactly should be characterised? This question is rarely addressed. Most discussions begin with an empirical study, often ad hoc, whose data are to be summarised and interpreted for a scientific publication. This is not objectionable, but problems arise when such individual study results are used in arguments for or against the use of a particular type of forensic examination (e.g., comparative toolmark examination) in legal proceedings. The latent assumption here is that the characteristics of the case at hand are somehow comparable to those of the validation study from which a specific error rate figure is derived. However, strictly speaking, we know that this is not the case. The (trace) items examined in the case at hand were generated under *un*controlled conditions, whereas examinations in validation studies are usually conducted under controlled laboratory conditions. And that is only the first of a cascade of assumptions.[2]

The deeper problem here is the flawed analogy between validation studies for traditional testing devices, such as those used in medical diagnostics, and forensic comparisons in the context of source inference (Biedermann, 2022). In medical diagnosis, the conditions under which the validation study is conducted – for example, the use of good-quality blood samples and a given

[2] In this context, see also Imwinkelried (2020) on the notion of "range of validation".

industrially produced examination kit or device – will be very similar to those under which the product is deployed for use in practice. This is why the performance *of a given testing device* measured during the validation study can provide useful information this performance in operational practice. This is not even remotely similar to comparative examinations in forensic science. In forensic examinations, the items to be analysed (of unknown source) are generated under *uncontrolled* conditions and vary wildly in type and quality. Furthermore, there is no such thing as a particular standardised examination device: rather, there are as many black boxes as there are human examiners. Therefore, the idea of using a single summary statistic drawn from an idiosyncratic empirical study to characterise the "validity" of a particular forensic discipline, method, or cohort of examiners, is implausible. At the very least, descriptively, such a statistic would, due to its generality, favour the poor examiner and penalise the above-average examiner.

Nevertheless, this does not mean that black box data are entirely useless. When broken down to the level of the individual examiner, they may reveal useful information, which takes us back to Luby and Kadane's (2025) divide-and-conquer approach and their recommendation to record data at the level of the individual examiners. In other words, we may inquire about questions such as "How assertive is *this* examiner?" (i.e., what proportion of responses other than 'inconclusive' has *this* examiner given?), even when the answer to such questions can only be based on data collected in examinations conducted under controlled conditions (i.e., a black box study), examiners may behave differently in casework than in controlled studies[3] (Dror, 2025) and examiners would need to agree to reveal their identity as a study participant.[4]

More generally, another important reason why black box data is not completely useless becomes apparent when we consider one of the primary reasons for invoking error rates in the first place: to help legal decision-makers assess the trustworthiness and informative value of the forensic results they hear. Without some sort of error rate data, even if it is imperfect at an individual level and, strictly speaking, not *directly* informative about the evidential value of testimony in the instant case, decision-makers will be unable to make even basic assessments of performance. For example, they will be unable to assess whether technique A is *generally* better than technique B (Koehler, 2008). The idea here that data provide a general anchor that can be refined further using additional, case-specific information about the examiner and the type, features, etc. of the examined item of evidence (Koehler, 2008).

The above critique does not suggest that the PCAST Report made an unsuitable recommendation when it stated that "[e]valuations of validity and reliability must (…) be based on 'black-box studies,' 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, 2015, p. 5–6). *At the time of publication*, this recommendation was suitable to achieve some progress quickly. However, the problem is that this recommendation has come to be seen as *the* standard to be pursued without further consideration. Indeed, a stream of publications now apply a black-box study blueprint across a variety of disciplines. This is not to deny the value of this type of research in getting at least something done, but it does not address the elephant in the room: the traditional forensic conclusion categories 'identification', 'exclusion' and 'inconclusive'. After all, the debate over how to score 'inconclusive' responses exists only because this reporting format was accepted

[3] This point has long been recognised and has been the main reason behind calls for *blind* proficiency testing in forensic science for several decades. For a recent commentary on this topic and further reading on this topic, see, for example, Dror (2025).

[4] For an example of how empirical study data might be used to characterise an examiner's performance alongside other factors, see Eldridge and Champod (2020).

in the first place. Therefore, engaging in black box studies and analysing the resulting data amounts to tacitly approving and perpetuating this reporting framework (Biedermann 2022).

Critics may argue that engaging with black box study data is akin to engaging with reality, and failing to assist the judiciary in addressing the current state of forensic reporting would be inadequate. In the short term, this may be a valid argument, but it raises the question of whether this is the best research perspective that academics can pursue. It is legitimate to ask whether it is the duty of academics to point out and address the root cause of the problem, and abandon the traditional reporting scheme of 'identification', 'exclusion' and 'inconclusive'. As noted by Morrison (2022) and others before, a complete alternative conclusion and reporting scheme is available. It is based on expressions of probative value on a continuous scale and includes metrics for evaluating performance. A core feature of this approach is focusing on the value of the observed features (and related measurements) of the compared forensic traces and abstaining from providing a direct opinion on propositions regarding the source of the compared items.

The call to abandon the traditional forensic reporting scheme in favour of one that provides an assessment of the discriminative capacity of the observed trace features is about more than just the pursuit of conceptual niceties. It involves breaking down the process of forensic comparison and evaluation of comparison results to its core components, and then reconstructing the task coherently as a problem inference, i.e. reasoning under uncertainty (Biedermann and Kotsoglou, 2021; Biedermann and Champod, 2025). It is important to clarify what this perspective entails: it focuses on evaluating the *selectivity* of the observed trace features, as this is the primary focus of comparative forensic examination. For example, a high-quality fingermark with twenty or more minutiae is clearly more selective (or discriminative) and hence informative than a poor-quality smeared fingermark and fewer than ten minutiae. The concept of feature selectivity aims to capture the discriminative capacity of observed trace features. Focusing on the traditional reporting scheme means to turn a blind eye on these distinctions. Why? Because, when focusing solely on whether an examiner reports 'identification', for example, and then inquiring into the value of such an expert utterance, based on black box study data, amounts to defining as the evidence the expert's utterance, *not* the observed trace features. This amounts to examiner diagnosticism (Biedermann 2022). In other words, examiner diagnosticism characterises the diagnostic value of the expert's utterance, rather than the value of the actual forensic trace. Examiner diagnosticism is neither a substitute nor a proxy for feature selectivity, since if an examiner utters 'identification' (or some other categorical conclusion), the underlying trace to which the testimony refers may be anything, from a trace of poor quality to one of high quality. Equating the diagnosticity of an expert's utterance with the selectivity of a forensic trace would be a gross mischaracterisation and illustrates the drawbacks of reductionism and oversimplification. Nevertheless, some jurisdictions permit and tolerate direct and categorical opinions by specialised witnesses on propositions. Furthermore, the lack of data to characterise feature selectivity in alternative reporting formats actually reinforces this practice.

As there is currently limited prospect of paradigmatic change, especially in the field of fingerprint examiners (Swofford et al., 2020), the approach by Luby and Kadane's (2025) can be expected to improve understanding of data from black box studies that use the traditional forensic reporting scheme as the relevant reference point. However, this should not distract us from the fact that the future of forensic science lies elsewhere: "The scientific reinvention of forensic science" (Koehler et al., 2023) involves abandoning unscientific reporting in the form of categorical and probabilistic opinions on propositions regarding the source of forensic traces.

This includes 'inconclusive' conclusions, not least because a complete and operational alternative framework is available (Morrison 2022).[5]

Acknowledgments

The author is grateful to Professor Koehler for the invitation to write this commentary and for his valuable comments.

Conflict of interest statement

The author declares that this commentary was written in the absence of any commercial or financial relationship that could be construed as a potential conflict of interest.

Bibliography

Arkes H.R. and Koehler J.J. 2021, Inconclusives and error rates in forensic science: a signal detection approach, *Law, Probability & Risk*, 20, 153–168.

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. and Champod C. 2025, 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*, 29, 140–160,

Biedermann A. and Kotsoglou K.N. 2021, Forensic science and the principle of excluded middle: "Inconclusive" decisions and the structure of error rate studies, *Forensic Science International: Synergy*, 3, 10047.

Curran J.M. 2013, Is forensic science the least bastion of resistance against statistics?, *Science & Justice*, 53, 251–252.

Dror I., Letter to the Editor – Black-box studies do not reflect decisions and errors in casework, *Journal of Forensic Sciences* (2025, in press).

Dror I.E. and Scurich N. 2020, (Mis)use of scientific measurements in forensic science, *Forensic Science International: Synergy*, 2, 333–338.

Eldridge H. and Champod C. 2020, Expert fingerprint examination, A primer on error rates, available at: https://zenodo.org/records/3734560 (last accessed June 14 2025).

Imwinkelried E.J. 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.

Koehler J.J. 2008, Fingerprint error rates and proficiency tests: what they are and why they matter, *Hastings Law Journal*, 59, 1077–1100.

[5] For an example in the field of forensic voice comparison, see Morrison et al. (2021). For examples in the field of fingerprint examination, see e.g. Neumann et al. (2012) and Swofford et al. (2018).

Koehler J.J., Mnookin J.L., Saks M.J. 2023, The scientific reinvention of forensic science, Proceedings of the National Academy of Science, 120, e2301840120.

Luby A. and Kadane J.B., A variance decomposition approach to inconclusives in forensic black box studies, *Law, Probability & Risk* (2025, in press).

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, 299–309.

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

Neumann C., Evett I.W., Skerrett J. 2012, Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm, *Journal of the Royal Statistical Society* (Series A), 175, 371–415.

President's Council of Advisors on Science and Technology (PCAST) 2016, Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods, Washington, D.C.

Swofford H., Cole S., King V. 2020, Mt. Everest–we are going to lose many: a survey on fingerprint examiners' opinions on probabilistic reporting, *Law, Probability & Risk*, 19, 255–291.

Swofford H., Lund S., Iyer H., Butler J., Soons J., Thompson R., Desiderio V., Jones J.P., Ramotowski R. 2024, Inconclusive decisions and error rates in forensic science, *Forensic Science International: Synergy*, 8, 100472.

Swofford H.J., Koertner A.J., Zemp F., Ausdemore M., Liu A., Salyards M.J. 2018, A method for the statistical interpretation of friction ridge skin impression evidence: Method development and validation, *Forensic Science International*, 287, 113–126.