Law, Probability and Risk, 2025, **24**, 1–30
https://doi.org/10.1093/lpr/mgaf004
Research Article

OXFORD

Evaluating reporting practices in fingerprint comparisons using information theory: five response categories are better than three

Andrew L. Cohen[1,*], Jeffrey J. Starns[1], Meredith Coon[2], Nada Aggadi[3], Thomas A. Busey[3]

[1]Department of Psychological and Brain Sciences, University of Massachusetts, Amherst, MA 01003, United States
[2]Aver, LLC, Washington, DC 20002, United States
[3]Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405, United States

*Corresponding author. Department of Psychological and Brain Sciences, 135 Hicks Way, Amherst, MA 01003, United States. Email: alc@umass.edu.

Abstract

Forensic fingerprint examiners communicate the results of their comparisons using a categorical decision scale that traditionally has only two or three categories. Newer scales with five categories have been proposed and, in some cases, adopted. In the present work, we model existing fingerprint comparison data using an approach that characterizes the amount of information provided by examiner decisions when they use a given scale. We demonstrate that the five-response-category scale produces an overall higher information gain, maintains high information gain even for risk-averse examiners, is more tolerant of non-optimal decision thresholds, and tends to encourage examiner decision thresholds that are very close to optimal, especially the critical decision threshold that determines an identification response. We further show that all of these advantages of the five-response-category scale are robust against variations in the base rates of a mated pair. In addition, there appear to be relatively few negative aspects of this scale, and larger scales are unlikely to produce marked improvements over five categories. The current work is an endorsement of efforts to shift to five-response-category decision scales in the friction ridge discipline.

Keywords: fingerprints; expected information gain; information theory; response categories; signal detection theory (SDT).

1. Introduction

Forensic fingerprint examination is a technique that provides information about the source of a latent impression deposited at a crime scene. Historically, examiner observations have been reported as categorical decisions that provide varying levels of support for or against the proposition that the latent print and a comparison print, for example, a print taken from a suspect, are from the same source. For example, an *identification* response indicates that the examiner believes the same finger made the two impressions and an *exclusion* response indicates that the examiner believes different fingers made the two impressions.

Although the specific response terminology has changed over time, fingerprint examination decisions have commonly been made on either a two-response-category scale (e.g., not identification/identification) or a three-response-category scale (e.g., exclusion/inconclusive/identification). More recently, standardization groups such as the Organization of Scientific Area Committees (OSAC 2018) and ANSI National Accreditation Board (ANAB) Friction Ridge Working Groups (ASAB 2024) have suggested a move to a five-response-category scale (e.g., exclusion/inconclusive with dissimilarities/inconclusive/inconclusive with similarities/identification). To date, relatively little work

Received: 18 June 2024. **Revised:** 3 May 2025. **Accepted:** 6 May 2025

has empirically addressed the consequences of expanded decision scales (Carter et al. 2020; Busey et al. 2022) and none have explored the theoretical value of expanded scales. The main goal of the current work is to determine the relative value of two-, three-, and five-response-category scales to the criminal justice system.

An identification response is of greatest interest to the legal system, as it potentially relates an individual to a crime scene. Erroneous identifications increase the odds of convicting an innocent person and are considered the most egregious type of error in most legal systems. As such, our discussion of different response scales acknowledges that minimizing false identifications is an important goal. However, one cannot assess the comprehensive value of an evidence-gathering procedure, such as fingerprint examinations, by evaluating a single response category (identifications) or a single potential consequence (convictions). For example, exclusion responses will sometimes have meaningful exculpatory value (e.g., if a clear latent print was taken from the murder weapon) and fingerprint evidence will sometimes be used to inform decisions other than a trial verdict (e.g., whether an arrest warrant is granted). Thus, researchers and policy makers need a measure of evidentiary value that considers all potential outcomes of a given fingerprint comparison procedure and assesses the fundamental purpose of fingerprint comparison, namely, to provide information about whether latent and comparison prints share the same source. Ideally, researchers will be able to specify procedures that achieve practical goals (e.g., minimizing false identifications) while simultaneously maintaining a high information value.

A key contribution of this work is to assess the value of the response scales using the theoretically grounded and highly influential measure, Expected Information Gain (EIG; e.g., Shannon 1948; Starns, Cohen, and Rotello 2023). Other common approaches to measuring the value of a response scale, such as ROC and AUC, are addressed in the General Discussion. When applied to fingerprint examination, EIG provides a measure of the amount of information that will typically be gained about the relationship between the latent and comparison prints. In particular, a high EIG means that an examiner's response is likely to provide substantial information regarding whether or not the prints are from the same source. Similarly, a low EIG means that the addition of an examiner's response is likely to add little information. Critically, EIG considers *all* possible outcomes of a response scale, for example, exclusion, inconclusive, and identification for a three-response-category scale, in a theoretically principled way. Furthermore, EIG directly assesses the amount of information gained by a consumer (e.g., investigators, judges, attorneys, juries, etc) about the critical question of interest—Are the latent and comparison prints from the same source?

EIG is well situated to contrast the information provided by different response scales, which we assess in two main ways. First, EIG is determined for empirical data collected from examiners using three- and five-response-category scales. A two-category response scale, derived from the three-category response scale responses, is also considered. To preview, there is a large increase in information gained when moving from a two- to a three-response-category scale, and a smaller increase when moving from a three- to a five-response-category scale.

Second, we use EIG to evaluate how the available response categories affect examiner decision thresholds and how these decision thresholds relate to the optimal decision thresholds. The decision provided by a fingerprint examiner is the result of a tradeoff between two objectives: first, increasing the likelihood of *correctly* expressing support for the proposition that two impressions are from the same source and, second, decreasing the likelihood of *incorrectly* expressing support for the proposition that two impressions are from the same source. Under common assumptions, these two goals are inversely related, as increasing the likelihood of the former decreases the likelihood of the latter, and vice versa.

This tradeoff is typically managed by adjusting decision thresholds (Thompson 2023). A decision threshold determines the amount of evidence required to produce a particular response, for example, a substantial amount of evidence that two prints are from the same source would reasonably be required to justify an identification response. Adjustments to the decision thresholds determine the balance between the two tradeoff objectives. For example, a risk-averse examiner requires more evidence to produce an identification response, which, in turn, reduces the likelihood of investigating an innocent person, but also increases the likelihood of exculpating a guilty person. In contrast, a risk-seeking examiner requires less evidence to produce an

identification response, which decreases the likelihood of exculpating a guilty person at the cost of increasing the likelihood of investigating an innocent person.[1]

Optimal decision thresholds perfectly balance these two tradeoffs. Here, we define optimal decision thresholds as thresholds that provide the most information about whether or not a print pair is from the same source. Thus, we define the optimal response policy for a given scale by determining the decision thresholds that maximize EIG. A model of the decision process is needed to determine the optimal decision thresholds. That is, a theoretical framework is needed in which the effect of moving the decision thresholds can be explored. We follow Mannering et al. (2021) and use a SDT model. Using a SDT model allows us to not only determine the relationship between optimal and empirically-determined examiner decision thresholds, but also allows us to establish whether expanded decision scales provide additional value. To preview the main conclusions: Empirical decision thresholds are close to optimal for the five-response-category scale, but risk averse for two- and three-response-category scales; both optimal and empirical EIG increase with the number of response categories, in particular, the empirical EIG for the five-response-category scale is close to optimal; the five-response-category condition is more robust to variations of thresholds; and the relative performance for the different response scales remains consistent across a wide range of base rates, that is, how often print pairs are from the same source.

The rest of the article proceeds as follows. First, we provide an overview of EIG and its role in fingerprint analysis. Second, we describe the SDT model of fingerprint decisions. Third, we explore how the number of response categories affects decision thresholds and how the decision thresholds used by examiners relate to the optimal decision thresholds. Fourth, we consider the effect of number of response categories and threshold placement on EIG. Finally, we consider the impact of base rates on our analyses.

2. EIG in fingerprint analysis

EIG is a measure of the overall value of an evidence-gathering procedure. EIG tells you how much information is likely to be gained about the true state of the world when all possible outcomes of the procedure are considered. According to EIG, an evidence-gathering procedure is valuable if it has a high potential to convince someone of what is actually true. EIG and related measures have been influential across many fields (Chaitin 1975; Benish 1999; Cover and Thomas 2006), including feature-based fingerprint analysis (Osterburg et al. 1977) and other forensic applications (Campbell et al. 2005; Ramos et al. 2007; Ramos and Gonzalez-Rodriguez 2008; Gong et al. 2015; Starns, Cohen, and Rotello 2023). This section provides a conceptual overview of the use of EIG to fingerprint examinations. More details about EIG, including computational formulas, are provided in Appendix A.

In the case of fingerprint examination, the true, unknown state of the world is the relationship between the latent and comparison prints. Whereas *mated* pairs of prints share a common source, *non-mated* pairs of prints come from different sources. The evidence-gathering procedure is a fingerprint examination. The outcome of a fingerprint examination is a judgment about the true state of the world. That is, the outcome is a statement about the degree of belief in the *common-source* and *different-source* hypotheses, that is, whether the prints are mated or not, respectively. These judgments are typically reported as categorical decisions, as discussed previously. For example, an identification decision expresses a strong belief in the common-source hypothesis.

The critical comparison underlying EIG is between the uncertainty of the true state of the world before and after the evidence-gathering procedure. EIG increases with the ability of the evidence-gathering procedure to increase certainty regarding the true state of the world. Consider a hypothetical laboratory where all samples are submitted by omniscient investigators, who always submit mated pairs. There is no uncertainty about the true state of the world, even before examination, because the submitted pairs are always mated. The examination procedure

[1] For ease of exposition, and because it is the most impactful decision, we focus on the decision threshold that determines an identification response. There is, however, a decision threshold between each pair of response categories.

cannot reduce uncertainty, regardless of the quality of the procedure, thus EIG is 0 even though the process is free of errors. Next, consider investigators that are known to submit 50% mated pairs and 50% non-mated pairs for examination. Before examination, the uncertainty about the true state of the world is at the maximum—Both outcomes are equally likely and there is no information about which outcome is true. In this scenario, the examination procedure has the possibility to reduce uncertainty. That is, a quality examination procedure can now provide information to the investigators, that, on average, reduces uncertainty of whether the prints are mated, leading to a higher EIG. In the scenarios we explore here,[2] the highest possible EIG, which would be produced by a combination of evenly divided samples submitted and a perfect examiner, is 1.00. In contrast, a poor examination procedure, say a faulty computer system that reports random decisions, provides no information about whether the prints are mated, thus, uncertainty is not reduced by the evidence-gathering procedure, again leading to an EIG of 0.

Because EIG is defined, in part, by the uncertainty of the true state of the world prior to the examination, it is necessary to determine the pre-examination uncertainty. As established by Starns, Cohen, and Rotello (2023), the evidentiary value of an information source is often best defined by assuming that no prior information is available; that is, by initially placing equal credibility in the common-source and different-source hypotheses. This assumption is certainly warranted for the data we consider here, which comes from an experiment in which examiners encountered half mated and half non-mated pairs. We further examine this assumption in a later section.

It is important to note that EIG does not measure the reduction in uncertainty due to a *single* response, but rather the *average* reduction in uncertainty across the entire range of possible responses. Thus, according to EIG, there are two ways to improve the fingerprint examination process. First, you can increase the likelihood of being able to make informative decisions, such as identifications and exclusions. One possible way to do this would be to increase the quality of the prints considered for examination. Poor-quality prints are more likely to produce inconclusive or erroneous results, which, in turn, leads to less informative decisions, reducing EIG. In contrast, high-quality prints allow for more informative decisions, such as identifications and exclusions, which increases EIG.[3] Second, you can increase the amount of information provided by the examination decisions. One way to do this is to select a response scale that more accurately conveys the examiner's assessment of the evidence to the investigator. We explore two factors that potentially affect the information value of the response scale in the current work: whether increasing the number of response categories supports increased information gain, and how inappropriate use of these response categories, for example, an overly cautious examiner, affects information gain.

Most evidence-gathering procedures produce an internal outcome that is typically assumed to be continuous in nature, for example, the amount of visual evidence for or against a mated pair. In the case of a categorical scale, decision thresholds are applied that partition this continuous space into discrete categories. This partitioning necessarily loses information, for example, from the categorical response alone, it is impossible to know whether an identification decision resulted from a value near the decision threshold (a borderline case) or far from the decision threshold (a conclusive case).

The location of the decision threshold also plays a major role in the information provided by the examiner. Taking the example of a two-response-category scale (e.g., identification or not identification), an extremely risk-averse examiner would require so much evidence to reach the identification threshold that identification responses are rarely produced and non-identification responses are very likely for both mated and non-mated pairs. As such, the examiner responses would provide little indication of which pairs truly share a common source. Thus, as discussed previously, the number of categories, and the choice of decision thresholds, can affect the amount of information provided by examination decisions. Because EIG quantifies the information provided by the examination procedure, we also use EIG to optimize the parameters of the

[2] That is, assuming equal likelihood of mated and non-mated pairs before examination.

[3] Note that this shift is due to an increase in the ability to make informative decisions, but the decisions have to remain informative. That is, simply shifting all decisions to the identification or exclusion categories by reducing the amount of evidence needed to make these decisions, in turn, reduces how much information these decisions provide, thus decreasing EIG.

system, that is, to determine the optimal number of decision categories and the location of the decision thresholds that define the decision categories.

In summary, we use EIG as a measure of how informative a fingerprint examination procedure is. More specifically, EIG measures how much an examination procedure reduces the uncertainty about whether two prints are mated. EIG considers the average information gain produced by the procedure; that is, EIG takes into account both how much information is gained by each decision category *and* how often these decision categories are produced. Whereas an EIG of 0 means that the examination procedure provides no additional information about whether the prints are mated, in the current context, an EIG of 1 means that the examination procedure is maximally informative. We use EIG to compare response scales of different sizes, as well as to determine the optimal location of the decision thresholds for a given scale. First, we need to introduce an application of SDT, as described next.

3. A Bayesian signal detection model of fingerprint identification

SDT models have been highly successful at accounting for and exploring psychological data in many domains, including perception (Haase, Theios, and Jenison 1999), memory (Banks 1970), and decision making (Phillips, Saks, and Peterson 2001). SDT models have also been previously applied to fingerprint analysis (e.g., Mannering et al. 2021; Smith and Neal 2021; Thompson 2023). This model assumes that a to-be-detected signal, in this case whether two impressions share a common source, is embedded in noise and the combined signal strength can be represented as a value along a single axis. While this axis is traditionally labeled signal or memory strength, here we might label this axis as "the amount of support for the same source hypothesis" or "the amount of perceived detail in agreement". The theoretical endpoints of this axis might be the most support imaginable for the different-source proposition and the most support imaginable for the same-source proposition. The model assumes that the comparison produces a value along this axis. This value is then compared against a set of decision thresholds to determine the term that is used to communicate the results of the comparison to the criminal justice partner (e.g., a judge, jury, defense attorney, or prosecutor). In the case of a five-response-category scale (e.g., exclusion, support for different source, inconclusive, support for same source, and identification), there are four thresholds, one between each pair of responses. If, for example, the value along the axis exceeds the support for same source/identification threshold, the examiner would report an identification decision.

Consider the SDT model of fingerprint analysis shown in Fig. 1. SDT models are generally used to determine the extent to which a proposition holds for a particular stimulus. For example, in the current context, a stimulus is a pair of latent and comparison prints and the two propositions are whether the two prints come from the same source or from different sources. The level of support for the same source proposition is represented by the x-axis in Fig. 1, where higher values indicate stronger support. Likewise, lower values indicate lower support for the same source proposition, or, equivalently, stronger support for the different source proposition.

A key assumption of SDT models is that, for a given latent-comparison print pair, the level of support for the proposition varies. That is, the strength of support for the same source proposition is a function of both the print pair and noise, which can be due to properties of the print pair or examiner. The noise is typically assumed to be normally distributed, and we make that assumption here. In Fig. 1, the dashed- and solid-line distributions are associated with non-mated and mated print pairs, respectively. Mated pairs tend to produce higher values because they will typically produce greater support for the same source proposition. Due to noise, however, non-mated pairs can provide some support for the same source proposition, and mated pairs can provide minimal support for the same source proposition, leading to overlapping distributions. The amount of overlap between these two distributions determines the *sensitivity* or *discriminability* of the examiner, that is, how well these non-mated and mated pairs can be discriminated. Lower distribution overlap indicates higher discriminability, and so non-mated and mated pairs are less likely to be confused. Note that the different natures of non-mated and mated pairs lead to different levels of uncertainty, that is, the non-mated and mated distributions

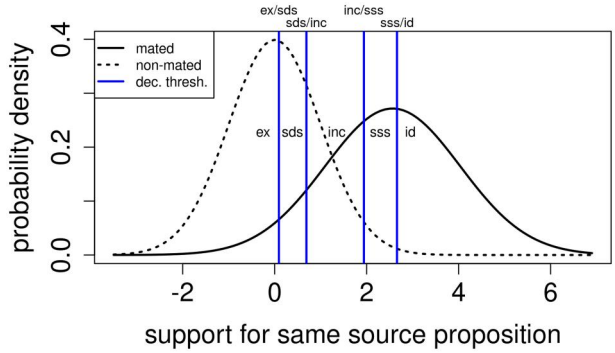

Figure 1. A signal detection model of fingerprint identification for the five-response-category condition. The dashed and solid curves represent distributions of support for the same source proposition when the fingerprints are non-mated and mated, respectively. The mean and standard deviation of the non-mated distribution is fixed at 0 and 1, respectively. The mean and standard deviation of the mated distribution are free to vary. The vertical lines are decision thresholds that separate the response categories. The response category labels are: exclusion (*ex*); support for different source (*sds*); inconclusive (*inc*); support for same source (*sss*); and identification (*id*).

have different standard deviations (see Fig. 1), which is a common assumption in signal detection models that is well supported by data (Wixted 2020).

The generated support for the same source proposition is then translated into a response or decision by comparing the level of support with a set of *decision thresholds*. Each decision threshold represents the level of support needed to make a particular response. In Fig. 1, there are four decision thresholds which separate the space into five response categories. Although the SDT model is agnostic regarding the category labels, we follow Busey et al. (2022) and assume the response categories are, from left to right, exclusion (*ex*), support for different source (*sds*), inconclusive (*inc*), support for same source (*sss*), and identification (*id*). We indicate specific decision thresholds by listing the two response categories they separate. For example, *ex*/*sds* would be the decision threshold that separates the *ex* and *sds* response categories. In a hypothetical sample, because it is lower than the *ex*/*sds* decision threshold, a level of support for the same source proposition of −2 would lead to an *ex* response (see Fig. 1). A level of support for the same source proposition of 2 would lead to a *sss* response because it is between the *inc*/*sss and sss*/*id* decision thresholds. Higher decision thresholds, for example, increasing the position of the *sss*/*id* decision threshold, have traditionally been described as more risk averse, because a higher amount of support is needed to generate decisions that indicate support for common source, for example, an identification decision. Similarly, lower decision thresholds are described as more risk-seeking.

This SDT model has $2 + (r-1)$ parameters: The mean, μ, and standard deviation, σ, of the mated pair distribution, and the $r-1$ decision thresholds, where r is the number of response categories. For two, three, and five response categories that means there are four, five, and seven parameters. Without loss of generality and to set the scale, the mean and standard deviation of the non-mated pair distribution are assumed to be 0 and 1, respectively. Thus, the support for same source values is in units of the standard deviation of the non-mated pair distribution.

Given a set of parameters, predicted model response proportions can then be determined. In Fig. 1, there are ten possible responses: *ex*, *sds*, *inc*, *sss*, and *id* for both non-mated and mated pairs. Within each stimulus type, non-mated and mated, the response proportions across the five response categories sum to 1. A response proportion is given by the area under the relevant curve between the appropriate decision thresholds. For example, in Fig. 1, given a non-mated pair, the probability of an *sds* response is given by the area under the dashed curve between the *ex*/*sds* and *sds*/*inc* decision thresholds. Given a mated pair, the probability of a *sss* response is given by the area under the solid line curve between the *inc*/*sss* and *sss*/*id* decision thresholds. All ten response proportions can be computed similarly.

Note that the analyses presented below do not rely on the SDT model being the "correct" model of fingerprint analysis. The SDT model is used as a device for varying task behavior, for

example, by changing the decision thresholds for fixed distributions, so that different outcomes can be compared. To that end, what is required is that the model produces results that reasonably mimic human performance under different risk averse or risk-seeking conditions. Indeed, the model can accurately reproduce human examiner performance for the current task (see Appendix Fig. B2). Further, the Gaussian distribution has a long history of application to human behavior because the central limit theorem naturally produces this distribution through the sum of several events (Thurstone 1927; Wixted 2020). Using a simple analogy, the SDT model serves a similar role in the current analysis as a crash test dummy when testing car safety. The dummies are clearly not human, but act in ways similar enough to humans in a crash context that car safety can be productively evaluated. Similar to changing the SDT model parameters, the height, weight, position, etc of the dummies can be varied to evaluate car safety under different circumstances. Thus, the SDT model of fingerprint analysis provides an extremely useful framework from which to evaluate differences across sets of fingerprint analysis response categories (Busey et al. 2022).

We implemented the SDT model in a Bayesian framework. One advantage of the Bayesian approach is that it allows for a natural expression of uncertainty in parameter estimates and evaluation values. That is, when applied to data, the model provides samples from a posterior distribution of parameter values (see Appendix Fig. B1). Statistics can then be applied to this distribution to determine, for example, the most likely range of parameter values. Throughout, all such credible intervals (CI) are defined using the 95% highest density interval (HDI), which contains the 95% most probable parameter values, given the data. For each sample from the posterior, we can also, for example, use the SDT model to make response proportion predictions (see Appendix Fig. B2) and to compute values such as EIG (see Fig. 2) and then determine HDIs.

Recall that the SDT model separates parameters related to sensitivity (the mean and standard deviation of the mated pair distribution) and parameters that determine the decision process (the decision thresholds). This separation allows us to explore different decision strategies for a fixed level of sensitivity. For example, for a given set of mated and non-mated pair distributions, we can determine the decision thresholds that maximize EIG—we refer to these as the *optimal* EIG values and decision thresholds. Similarly, for a given set of mated pair distribution parameters, we can determine the *maximum* possible EIG, that is, the EIG that would be produced if the fingerprint examiner were able to perfectly access and use the full underlying non-mated and mated distributions to determine whether the fingerprint pair was derived from the same source. HDIs can be determined for these values as described previously.

The model was applied to the data from Busey et al. (2022). In this experiment, examiners rated support for the same source proposition with either 3 (*ex*, *inc*, *id*) or 5 (*ex*, *sds*, *inc*, *sss*, *id*)

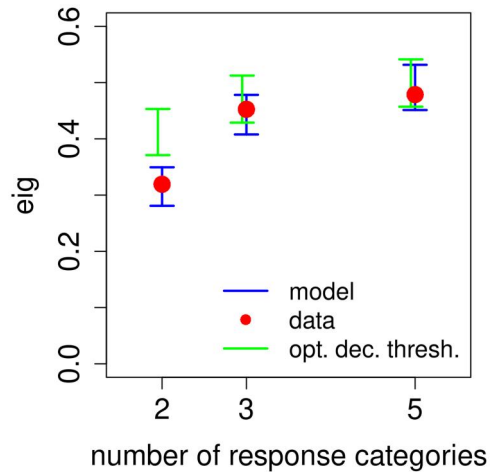

Figure 2. EIG as a function of the number of response categories. CIs are 95% HDIs. Opt. dec. thresh. is optimal decision threshold.

response categories. We make the assumption that the underlying sensitivity of examiners, that is, the ability of examiners to distinguish mated and non-mated pairs, is not affected by the available response categories which was supported by prior modeling in (Carter et al. 2020; Busey et al. 2022). Thus, we assume that both the three- and five-response-category conditions share a common μ_m and σ_m. The two decision thresholds in the three-response-category condition and the four decision thresholds in the five-response-category condition were allowed to vary independently. Thus, there are a total of eight parameters to account for 16 data points (non-mated and mated pairs for *ex*, *inc*, *id* and *ex*, *sds*, *inc*, *sss*, *id* response categories). In the analysis, we will also explore a two-response-category case (*not-id*, *id*), which was formed from the three-response-category data by combining the *ex* and *inc* response categories into a single *non-id* response category. The two-response-category case did not directly affect parameter estimation, and, because it was derived from the three-response-category case, shares a common decision threshold parameter (i.e., *not-id/id* = *inc/id*) to reflect the fact that this threshold will not change when the *ex* and *inc* categories are collapsed into *not-id*.

Further details for the analyses described in this section can be found in Appendix B.

4. Response categories and decision thresholds

First, we use the SDT model to explore how the available response categories affect decision thresholds and how the decision thresholds used by examiners relate to the optimal decision thresholds given the underlying non-mated and mated distributions and the response scale. The decisions produced using the optimal decision thresholds provide the most information regarding whether a pair is mated or not. The optimal decisions thresholds will not eliminate all errors, but rather they maximize EIG.

Figure 3 provides the results of the Bayesian SDT model for the two-, three-, and five-response-category conditions. As discussed previously, because we assume constant sensitivity across conditions, both the non-mated (dashed line) and mated (solid line) distributions are shared across conditions. The vertical blue lines, which we call the model decision thresholds, show the median decision thresholds that best account for the examiner data in each condition. The decision threshold that defines the identification region is very similar across conditions.[4] That is, the amount of evidence examiners required to make an identification only slightly varies based on the number of response categories. In particular, the *sss/id* decision threshold is only marginally more risk averse than the *not-id/id* and *inc/id* decision thresholds, as the 95% HDI of the difference is (0.02, 0.46).

The vertical green lines show the optimal decision thresholds. Compare the blue and green decision thresholds. In the two-response-category condition, examiners are significantly more risk averse than optimal, that is, the model decision threshold is significantly higher than the optimal threshold. The inset plot shows the median (circle) and 95% HDI (CI) of the difference between the model and optimal decision thresholds. Where the HDI does not overlap zero there is strong support that the optimal and model-based decision thresholds differ. In the three-response-category condition, examiners are less likely to make both exclusion and identification responses than is optimal. This result also shows up in the five-response-category condition, albeit to a slightly lesser degree. Examiner decision thresholds that define the exclusion and identification categories are much closer to the optimal thresholds in this condition, and in most cases the HDIs include zero.

In summary, regardless of the number of response categories, examiners set the decision threshold that defines an identification at a level similar to the optimal decision threshold that maximizes EIG in the five-response-category condition. The location of the examiner decision thresholds in the two- and three-response-category conditions is risk averse relative to the optimal decision thresholds. In the three- and five-response-category conditions, examiners are over-utilizing the inconclusive category. In the five-response-category condition, the examiner-based decision thresholds that define the exclusion and identification response regions are relatively well-calibrated.

[4] Recall that the *not-id/id* decision threshold for the two-response-category condition and the *inc/id* decision threshold for the three-response-category condition are identical.

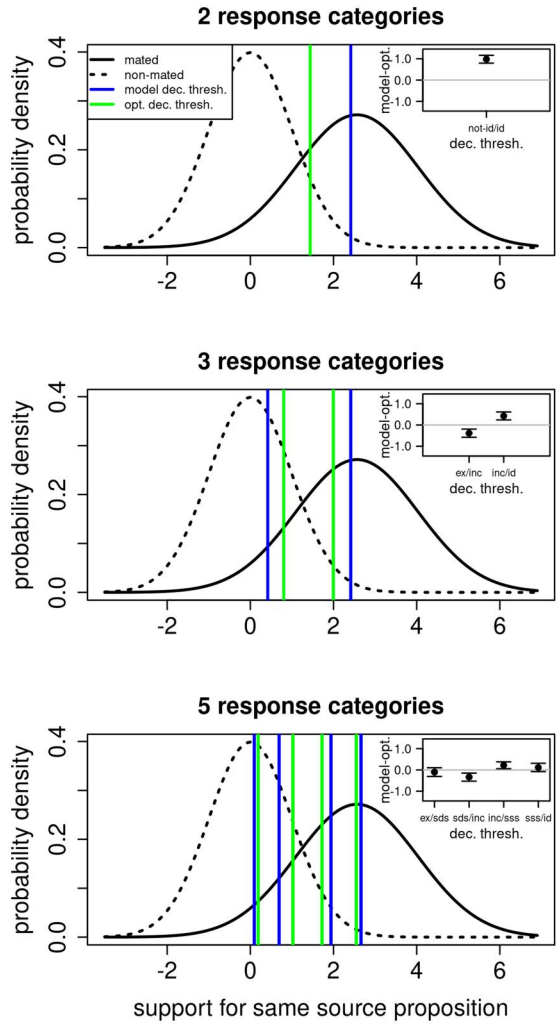

Figure 3. Results of the Bayesian SDT model for the two-, three-, and five-response-category conditions. The median posterior mated distribution and decision thresholds are shown. The insets show the median (circle) and 95% HDIs (CI) of the difference between the model and optimal decision thresholds. Dec. thresh. is decision threshold. Opt. is optimal. *ex* is the exclusion response category; *sds* is support for different source; *inc* is inconclusive; *sss* is support for same source; *id* is identification; and *non-id* combines the exclusion and inconclusive response categories from the three-response-category condition. For example, *ex/inc* is the decision threshold separating the exclusion and inconclusive response categories.

5. Response categories, decision thresholds, and EIG

Next, we consider the effect of number of response categories and threshold placement on EIG.

Figure 2 shows EIG as a function of the number of response categories. The red points are the EIG values determined directly from the response proportions in the examiner data. The blue confidence intervals show the 95% HDIs on EIG for each response category condition. Importantly, EIG increases with number of response categories. That is, as the number of response categories increases, examiners provided more information about whether the fingerprints are from the same source, albeit with decreasing gains between three and five response categories relative to the large gain from two to three response categories.

Readers might wonder what constitutes a meaningful change in EIG. This value varies across contexts, of course, but it is useful to consider some benchmarks in a simplified scenario in which EIG maps directly to a more familiar measure, proportion correct. Specifically, we consider the simplified scenario with two response categories (e.g., *not-id* and *id*) with the same level of accuracy for both outcomes; for example, if the proportion correct is 75%, then we assume for the purpose of demonstration that 75% of identification responses are made to mated pairs and 75% of non-identification responses are made to non-mated pairs. In this scenario, an EIG of 0.30 corresponds to a percent correct of about 81%, an EIG of 0.40 corresponds to about 85%, and an EIG of 0.50 corresponds to about 89%. Thus, across the range of observed values for this sample, an EIG change of 0.10 maps onto a proportion correct change of about 0.04. We hope that this reference point helps readers assess our reported EIG differences, for example, if a reader thinks that a factor that increases examiner accuracy by 2 percentage points is meaningful given the large number of cases that involve fingerprint evidence, then that reader should consider a 0.04–0.05 difference in EIG to be meaningful as well in the current context.

The green CIs show the 95% HDIs on EIG when using the optimal decision thresholds for each response category condition (see Fig. 1). Again, EIG increases with the number of response categories. This result is expected, as increasing response categories allows the model to provide more fine-grained information about the degree of support for the same source proposition. Of note, however, the difference between the optimal and model EIG values decreases as the number of response categories increases. That is, whereas examiners in the two-response-category condition provide far less information than is possible with optimal decision thresholds, examiners in the five-response-category condition are providing almost as much information as a model that uses optimal decision thresholds.

We can also consider how EIG changes as decision thresholds deviate from optimal, that is, how much information is lost when the decision threshold is either too risk-seeking or too risk averse. Although changes to any set of decision thresholds can be examined, because identifications are of particular interest, we focus on the decision threshold that determines the identification region (*not-id*/*id*, *inc*/*id*, or *sss*/*id* for two-, three-, and five-response categories, respectively). In particular, for each response category condition, we systematically varied the identification decision threshold from optimal, holding all other decision thresholds at optimal, to determine the effect on EIG (see Fig. 1 for optimal decision threshold positions). Note that EIG is still being computed across all response categories, but we are simply varying the identification decision threshold to see the effects of non-optimal identification decision thresholds on EIG.

The results are shown in Fig. 4. The *x*-axis represents the difference between the identification decision threshold and optimal decision threshold. A value of 0 on the x-axis indicates that the optimal decision threshold is used. Negative values along the x-axis indicate risk-seeking identification decision thresholds and positive values indicate risk-averse identification thresholds.

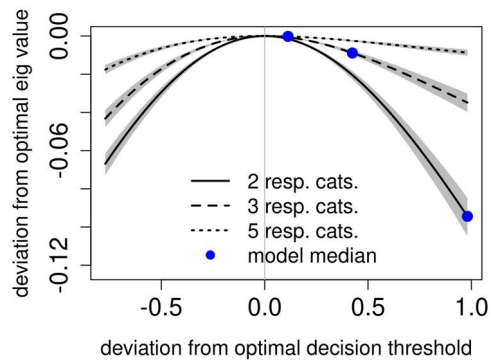

Figure 4. How EIG is affected by moving the identification decision threshold from the optimal position. All other decision thresholds are set at the optimal position. Resp. cats. is response category condition. The dots show the median results when applying the SDT model to examiner data. The shaded region shows 95% HDIs.

The key point is that the five-response-category condition is far more robust to variations in decision threshold position than the two- and three-response-category conditions. That is, whereas deviations from optimal significantly decrease EIG in the two- and three-response-category condition, EIG is relatively constant across a wide range of decision threshold positions in the five-response-category condition. In other words, moving the identification decision threshold from optimal has more of a deleterious effect on EIG if there are fewer other response categories. For an extreme example, in the two-response-category condition, moving the *not-id*/*id* decision threshold far to the right means that almost every decision will be *not-id* and so the examiner response provides almost no information about the prints. In contrast, in the five-response-category condition, moving the *sss*/*id* decision threshold to the right only affects the *sss* and *id* responses, effectively leaving the *ex*, *sds*, and *inc* response proportions unchanged and effectively turning the five-response-category scale into a 4-response-category scale (*ex*, *sds*, *inc*, *sss*). In short, additional response categories buffer EIG loss against variations in decision thresholds.

The blue dots in Fig. 4 show the results using the identification threshold that best fit the examiner data.[5] As discussed previously, the estimated identification thresholds move closer to optimal with increasing response categories. Coupled with the prior result that the five-response-category condition is more robust to variations of thresholds, less information is lost in the five-response-category condition compared to the three- and, especially, the two-response-category conditions.

Recall that, for fixed non-mated and mated distributions, the possible information gained increases with the number of response categories. Indeed, EIG is theoretically maximized by use of the entire underlying distribution, which is essentially using infinite response categories. Figure 5 shows the difference between the EIG achieved by examiners and this theoretically maximum EIG for each response category condition, assuming the estimated underlying distributions described previously. As the number of response categories increases, this difference decreases. Although the difference between the model and optimal EIG for the three- and five-response-category conditions was relatively similar, as shown in Fig. 2, it is clear from Fig. 5 that the five-response-category condition produces behavior that much more closely approaches the theoretical maximum. Thus, scales with even more response categories are likely to produce only very slight gains over a five-response-category scale.

A common way to visualize performance in signal-detection tasks is the receiver operating characteristic (ROC), which plots hit rates against false alarm rates across the range of decision thresholds. In the current context, we focus on identification rates, that is, how often the comparison print is identified as coming from the same source as the latent print. Recall that the identification rates for mated (hits) and non-mated (false alarms) pairs are given by the area under the mated and non-mated distributions from the SDT model, respectively, above the

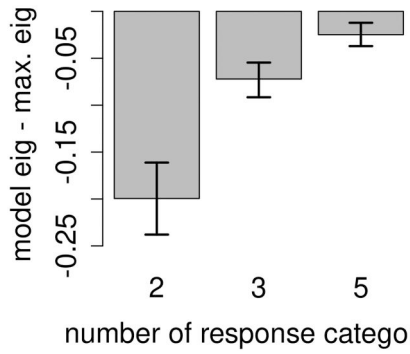

Figure 5. Difference between model EIG and maximum EIG (max. eig), using the underlying non-mated and mated distributions, as a function of the number of response categories. Bars closer to zero indicate a smaller difference between achieved and maximum EIG. Bars are medians and CIs are 95% HDIs.

[5] We used the median of the posterior distribution as a point estimate.

identification decision threshold (*not-id*/*id*, *inc*/*id*, *sss*/*id* in the two-, three-, and five-response-category conditions, respectively; Fig. 1).

These rates are defined for a given identification decision threshold. To create the ROC, the identification decision threshold is varied across a range, typically across the entire scale, and the corresponding true and false identification rates are plotted against each other for each decision threshold. With multiple response categories, however, varying the decision threshold across the full range is not feasible, as it doesn't make sense, for example, to move the *sss*/*id* decision threshold below the *ind*/*sss* decision threshold. We therefore vary the identification decision threshold across a restricted range. We start the identification decision threshold range at optimal (i.e., the threshold that maximizes EIG; Fig. 1) and then explore the effect of making this threshold increasingly risk-averse, that is, moving it to the right. Because the range is restricted, the ROC will not reach the point (1,1) as normally would occur if the range were unrestricted.

The ROC curves for the two-, three-, and five-response-category conditions are shown in Fig. 6. The rightmost point of each curve corresponds to the optimal *id* decision threshold, which will produce both the highest true and false *id* rates in the decision threshold range. The leftmost point of each curve corresponds to an extremely risk-averse *id* decision threshold, which would lead to near zero identification rates of both mated and non-mated pairs. Because they are generated from the same underlying SDT model distributions, these curves fall on top of each other. What varies across response category condition is the range of the curve. Because the optimal *id* threshold becomes more risk averse as the number of response categories increases (Fig. 1), the maximum true and false *id* rates decrease with number of response categories. Note that, whereas for two- and three-response-category conditions the false *id* rate is prohibitively high at the optimal decision threshold, for five-response categories, examiners achieve both a low false *id* rate and a close to optimal EIG.[6] In addition, in the five-response-category condition, false *id* rates can be further reduced without significant information loss.

For each *id* decision threshold, with all other decision thresholds set at the optimal values shown in Fig. 3, we can also compute the associated EIG value for the SDT model. The EIG value for each *id* decision threshold in Fig. 6 is indicated by brightness, with brighter points corresponding to higher EIG. This figure then provides us with another way to visualize the effect on EIG of encouraging a more risk-averse decision strategy (also see Fig. 5). In particular, as the *id* decision threshold moves to the right to become more risk averse, relative to the optimal *id* decision threshold, EIG decreases. Importantly, although the effect of changing the *id* decision threshold is severe for the two-response-category condition (i.e., points get much darker as the hit and false-alarm rates approach zero), and noticeable for the three-response-category condition, it is minimal for the five-response-category condition. That is, the five-response-category condition is robust to variation in *id* thresholds, allowing examiners to meet the joint goals of having very low false *id* rates and producing decisions with high information value.

6. The effect of base rate on EIG

To this point, we have assumed that the fingerprint comparison process begins with no information about whether a latent-comparison pair is mated or non-mated. Starting with no prior information is the only justifiable choice if the actual base rate of mated pairs is 0.50, that is, if 50% of the fingerprint pairs considered by examiners are mated. This assumption is warranted for the analyses reported here, because the base rate in the experiment that generated the analyzed data was 50% (Busey et al. 2022).[7] When searching impressions against a database, it is estimated that 22% of impressions result in possible identifications (Gardner, Neuman, and Kelley 2021), which suggests an overall base rate below 0.50. It should be noted, however, that base rates can vary dramatically across different scenarios, for example, comparisons stemming from a database search compared to comparisons to a suspect identified from other evidence. Indeed, base rates are likely to vary widely for different types of crime and across jurisdictions.

[6] 95% HDIs for false identification rates at the right-most points of the ROCs are (0.054, 0.101), (0.015, 0.032), and (0.003, 0.008) for the 2-, 3-, and 5-response-category conditions, respectively.

[7] Recall, however, that only "of-value" trials were analyzed, which lead to mated pair base rates of 0.511 and 0.513 in the 3- and 5-response-category conditions, respectively.

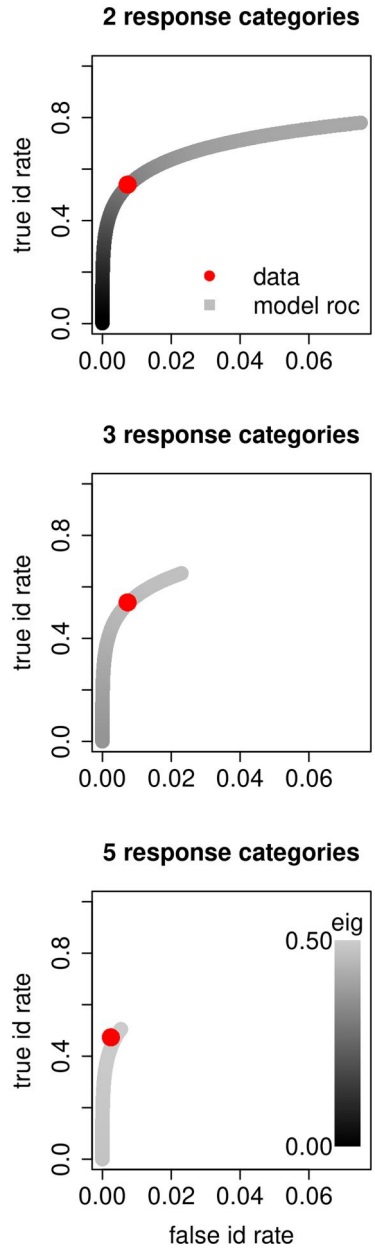

Figure 6. ROC curves based on *id* response proportions for the two-, three-, and five-response-category conditions. To create these curves, a restricted range was used for the *id* decision threshold—from the optimal *id* decision threshold and higher. The shade of each point on the ROC indicates the EIG associated with that *id* decision threshold. The red points indicate the true and false *id* rates in the data.

In this section, we consider whether our conclusions regarding number of response categories are robust to changes in base rates.

Consider Fig. 7, which illustrates how EIG changes with base rate for the two-, three-, and five-response-category scales examined previously. These EIG values were generated from the proportion of category responses from the observed examiner data by varying the proportion of mated pairs that were assumed to be shown to examiners. For example, for the three-response-category

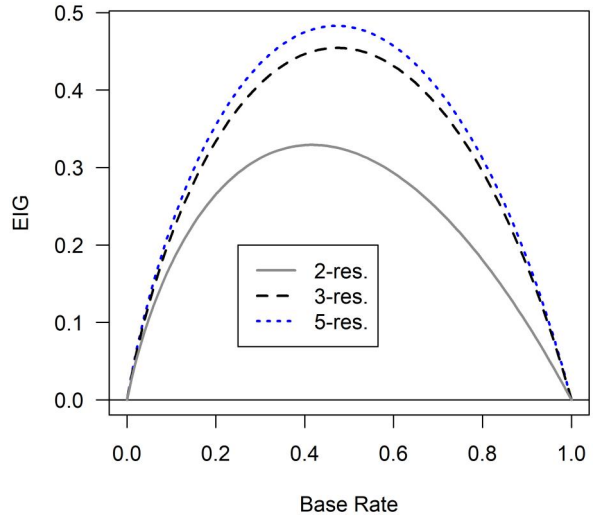

Figure 7. EIG for the observed response distributions for two-, three-, and five-response-category scales across all possible base rates of mated pairs.

condition, the probabilities of an exclusion, inconclusive, and identification response to mated pairs remained fixed at 0.068, 0.387, and 0.552 and the same probabilities to non-mated pairs remained fixed at 0.667, 0.327, 0.008 (see Table A1[8]), respectively, and EIG was computed using a range of the base rates of mated pairs that varied from 0 to 1 (see Appendix A for computation details).

Critically, although EIG necessarily changes across different base rates, the relative performance for the different response scales remains consistent across a wide range of base rates. That is, the three-response-category scale achieves a sizable information advantage over the two-response-category scale, with a smaller additional advantage when moving from the three- to five-response-category scale. Although this conclusion holds across the entire range, it is especially clear for mated-pair base rates between approximately 10% and 95% and strongest near 50%. Thus, the advantage of the five-response-category scale is robust to different assumptions about base rate. Figure 7 also shows that the EIG for any evidence-gathering procedure converges to zero as the base rate approaches 0 or 1. This pattern emerges because the prior distribution becomes extremely informative for very low or very high base rates, leaving little room to gain information from the examiner decision.[9]

Consideration of base rates may also affect other analyses. For example, recall that the five-response-category scale achieves a close-to-optimal EIG, even when the examiner does not make optimal use of the response categories, that is, even when the examiner decision thresholds deviate from the thresholds that maximize EIG. This property is important because it demonstrates that the five-response-category scale can be used to simultaneously produce informative examiner decisions and curtail false identifications with the use of identification thresholds that are far above the EIG-maximizing value. We evaluated whether this conclusion is also robust to changes in the assumed base rate. That is, for base rates between 0 and 1, we compared the EIG values when using the SDT-model thresholds estimated from examiner decisions and the SDT-model thresholds that optimize EIG.[10] This procedure was repeated for the two-, three-, and five-response-category scales.

[8] For example, $p(\textit{exclusion} \mid \textit{non-mated}) = p(\textit{exclusion} \cap \textit{non-mated})/p(\textit{non-mated}) = 0.326/0.489 = 0.667$. See Table A1.

[9] Extremely low or high base rates are undesirable for other reasons. Perhaps the most troubling issue is that, for extreme base rates, the reliability of an examiner response is often a misleading signal to the risk of making an error. For example, say that examiners are 100 times more likely to make an identification for mated than non-mated pairs. This value encourages a high level of trust in identifications as evidence for a common-source judgment. But if the base rate is 1%, then about half of identifications will be to non-mated pairs. Given the ubiquity of the fallacy of the transposed conditional, such a scenario sets the stage for egregious misuses of evidence. Moreover, as base rates become more extreme, the cost–benefit ratio of fingerprint examination decreases.

[10] The optimal thresholds change when the base rate changes, which was incorporated into the analysis.

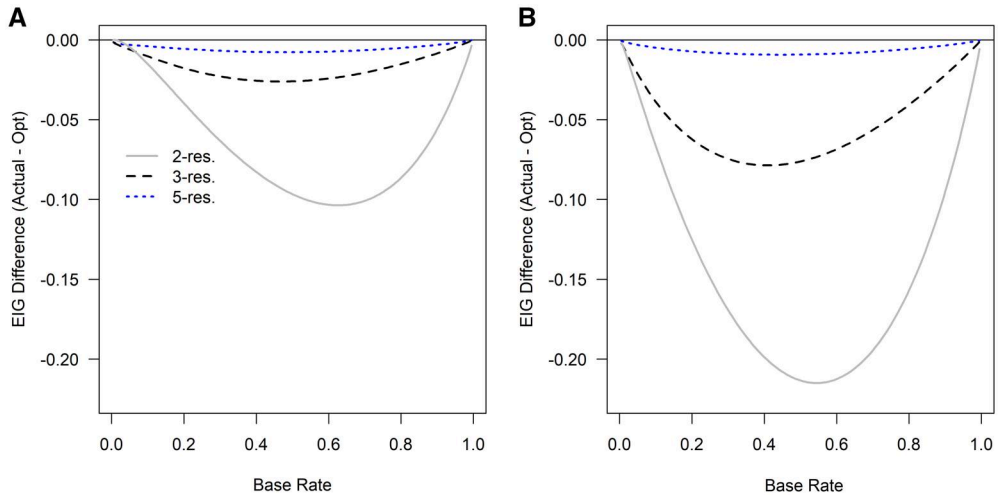

Figure 8. Reduction in EIG incurred by using a set of actual SDT-model thresholds (actual) relative to the thresholds that optimize EIG across a range of mated-pair base rates (opt). In Panel A, the actual SDT-model thresholds are set to the values that best fit the examiner responses. In Panel B, the identification threshold is adjusted to simulate examiners who will accept only a 0.001 chance of making a false identification, with all other thresholds set to the values that best fit the examiner responses. Res. is response categories.

Across base rates, Fig. 8 shows the reduction of EIG when the examiner thresholds are used, relative to the use of the optimal thresholds. Whereas a value of 0 means that the examiner thresholds achieve an EIG that is as high as the maximum EIG that could be produced by any set of thresholds, a negative value means that using the examiner thresholds resulted in a loss of information. Figure 8A contrasts EIG from the optimal thresholds and EIG from the thresholds that best fit the examiner data (Fig. 3). The five-response-category scale was more robust to deviations from the optimal thresholds (i.e., had an information loss closer to 0), and this result held across a wide range of base rates. The two-response scale incurred substantial information loss across a wide range of base rates, and the three-response scale was slightly worse than the five-response scale.

Interestingly, the information advantage of the five-response-category scale increases if the examiner identification threshold becomes more risk averse, as shown in Fig. 8B. The only change from Panel A to B is that the identification criterion was moved up to the value needed to achieve a very low false-identification rate (0.001). This shift to risk-averse responding has little impact on information loss for the five-response scale, which continues to produce EIG values that are close to the value achieved with optimal thresholds (information loss near 0). In contrast, information loss gets substantially worse for the two- and three-response-category scales when responding is very risk averse, demonstrated by the bigger "dips" in Panel B than Panel A. For example, with the thresholds from the examiner data and a 50% mated-pair base rate, the information loss for the two-, three-, and five-response-category scales is −0.10, −0.03, and −0.01, respectively. If the identification threshold is adjusted to achieve a 1 in 1000 false identification rate, however, information loss moves to −0.21, −0.08, and −0.01, respectively. Thus, whereas the two- and three-response-category scales incur substantial information loss if an extremely low false ID rate is enforced, the five-response-category scale maintains an information value near optimal.

7. Discussion

Some standards development organizations have recommended an expanded decision scale for reporting forensic fingerprint examination decisions (OSAC 2018; ASAB 2024), but relatively little work has empirically addressed the consequences of such a change. The main goal of the current work is to determine the relative value of two-, three-, and five-response-category scales.

A key contribution is to assess response scale value using EIG (e.g., Shannon 1948; Starns, Cohen, and Rotello 2023) in conjunction with a SDT model of fingerprint examination. EIG, when combined with SDT, acts as a form of alternative universe simulator: What would happen if we changed forensic procedures, in this case, the response scale, and how might differences across examiners or labs affect the performance of the entire system? When applied to fingerprint examination, EIG measures how much the fingerprint examination procedure, including use of the response scale, reduces the uncertainty regarding whether two prints are mated. We assessed empirical results from examiners when they used different response scales, and situated those results relative to the optimal results defined by the SDT model thresholds that maximized EIG.

The results demonstrate a clear superiority of a five-response-category scale across various metrics. In particular, the five-response-category scale produces an overall higher information gain, maintains high information gain even for risk-averse examiners, is more tolerant of non-optimal decision thresholds, and tends to encourage examiner decision thresholds that are very close to optimal, especially the critical *sss/id* decision threshold. All of these patterns are robust against variations in the base rates of a mated pair. In addition, there appear to be relatively few negative aspects of this scale, and scales with more than five-response categories are unlikely to produce marked improvements.

The current work clearly shows that the five-response-category scale should be preferred over the two- and three-response-category scales. For an examiner who appropriately maps the underlying available information to the response scale, performance would continue to improve with the number of response categories beyond 5. Indeed, with direct accurate access to the underlying evidence distributions (e.g., Fig. 1), a continuous response scale would maximize information gain. Such approaches, however, have their own strengths and weaknesses (Neumann, Evett, and Skerrett 2012; Swofford et al. 2018; Busey and Coon 2023). In practice, larger response scales are more difficult to use (Benjamin, Tullis, and Lee 2013) and require both examiners and consumers to correctly interpret the underlying scale. Notably, in absolute identification tasks, inconsistent scale use progressively degrades information value as the number of response categories increases over 5 even for simple attributes such as the loudness of tones (Garner 1953).

Larger scales are also more likely to promote inconsistency across examiners. As the number of potential responses increases, it becomes more difficult for different examiners to adopt the same interpretation for each response category, which can lead to costs similar to those of decreasing the ability of examiners to compare prints (Benjamin, Tullis, and Lee 2013). Increasing the number of response categories also exacerbates the already difficult problem of encouraging triers-of-fact to adopt appropriate interpretations (Martire, Kemp, and Newell 2013). Starns, Cohen, and Rotello (2023) demonstrated how EIG can be used to measure the cost of such misinterpretations. For example, EIG can be used to quantify the cost of incorrectly interpreting a support for same source decision that actually provides 4:1 evidence in favor of a mated pair as providing 20:1 evidence in favor of a mated pair.

Even if the response scale could be used and interpreted perfectly by both examiners and triers-of-fact, the current analysis shows that, in practice, each additional decision category adds less to the information gain. That is, the addition of response categories comes with diminishing returns. The five-response-category scale strikes a good balance as it is both easy to use and produces close-to-optimal information value.

As discussed previously, ROC plots and AUC are common methods for the visualization and measurement of performance in forensic science signal-detection tasks (e.g., Mickes, Flowe, and Wixted 2012; Albright 2022). Both EIG and AUC are measures of classification performance, and, in some circumstances, are nearly interchangeable. In particular, EIG and AUC will be nearly interchangeable to the extent that both measures are applied to the same evidence distributions, the distributions are evaluated as a stand-alone source of evidence (i.e., no other information is used), and the levels of evidence supported by examiner decisions are properly interpreted. We chose to use EIG given its potential to explore important factors such as the inclusion of other sources of information, for example, uneven base rates, and, as discussed

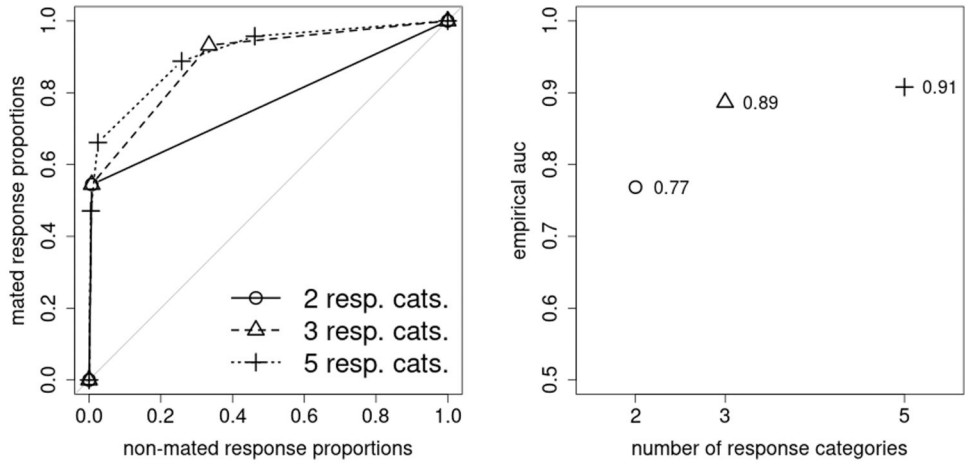

Figure 9. Left: Empirical ROCs for the two-, three-, and five-response-category conditions. Right: Empirical AUC values derived from the empirical ROCs of the two-, three-, and five-response-category conditions.

previously, the cost of scale misinterpretation. That said, because the current analyses met these criteria, AUC should support the same conclusions.

We therefore also provide ROCs and AUCs for the different response category conditions. In the previous analyses, EIG was used to measure performance supported by examiner *responses*. The corresponding ROC is called an *empirical ROC* (Wixted and Mickes 2018), and it is instantiated as a piecewise-linear function that plots observed response proportions for mated prints against observed response proportions for non-mated prints. An empirical AUC is the area under an empirical ROC, and provides a measure of how often an examiner decision could be used to successfully distinguish a mated and non-mated pair (Wixted and Mickes 2018). As the ROC moves into the upper-left corner, both performance and AUC increase. Figure 9 provides the empirical ROCs (left panel) and associated AUC values (right panel) for the two-, three-, and five-response-category conditions. The qualitative results are identical to that of the EIG analysis. In particular, AUC increases with the number of response categories, but with diminishing returns. Thus, in this case, discriminability tracks information gain.

The EIG analysis reported here is of clear and direct use to policymakers when the goal is to design a fingerprint analysis procedure that maximizes the information gained about whether a pair of prints are mated or not. EIG can be used to evaluate other questions of interest to policy makers, such as the interaction of multiple sources of evidence (e.g., Zane et al. 2025) and the effects of misinterpreting evidence (e.g., under- versus over-confidence; Starns, Cohen, and Rotello 2023). EIG is a summary measure across multiple cases, making it well suited to guiding policy decisions that will apply in multiple investigations. Triers-of-fact need to consider the evidence associated with individual cases, so they need to consider the accuracy of a specific examiner decision in addition to, or instead of, EIG.

Thus, in Table 1, we present the accuracy of different response categories for the two-, three-, and five-response-category conditions. For a given response category, accuracy is defined as # correct/(# correct + # incorrect) (see Confidence Accuracy Characteristic, CAC; Mickes 2015; Wixted and Wells 2017). For example, in the five-response-category condition, there were 210 and 2 identification responses to mated and non-mated pairs, respectively, which produces an accuracy measure of $210/(210+2) = 0.991$. Because there is no associated correct response, inconclusive responses are not included in this analysis. Exclusion (*ex*) and identification (*id*) responses are provided for all response category conditions. Although they cannot be compared across condition, accuracies for "support for different source" (*sds*) and "support for same source" (*sss*) are provided for the five-response-category condition.

Exclusion response accuracy shows the same pattern as EIG and AUC, that is, a sharp increase from two- to three-response categories and a significantly smaller increase from three- to five-response categories. Identification response accuracy is somewhat different. Because they were

Table 1. Accuracy of different response categories for the two-, three-, and five-response-category conditions.

	Response categories			
# of response categories	ex	sds	sss	id
2	0.676	–	–	0.987
3	0.904	–	–	0.987
5	0.920	0.728	0.914	0.991

Note. The response category labels are: exclusion (ex); support for different source (sds); support for same source (sss); and identification (id). Because there is no associated correct response, inconclusive responses are not included.

generated from the same data, there is no difference between two- and three-response-category identification accuracy. Consistent with EIG and AUC, however, there is only a slight increase in accuracy from three- to five-response categories. Although somewhat limited, this analysis suggests that accurate use of a given response category increases with the number of response categories, but with diminishing returns. These results also suggest that the evidence for a mated pair provided by "positive" responses (*id* and *sss*) is stronger than the evidence for a non-mated pair provided by "negative" responses (*ex* and *sds*).

Recall that the two-response-category condition data were produced by re-assigning inconclusive responses in the three-response-category condition as exclusions. Thus, the id decision thresholds in the two- and three-response-category conditions are identical. It is possible that examiners in an independent two-response-category condition would have used a different threshold, which, in turn, would affect EIG. We ask whether the relative poor performance of the two-response-category condition is due to this constraint, that is, whether there exists a two-response-category *ex/id* decision threshold that increases EIG to that of the three-response-category condition. To answer this question, we varied the *ex/id* decision threshold in the two-response-category condition across a wide range and determined EIG for each decision threshold. We assume the same underlying distributions from Fig. 3.[11]

The results are provided in Fig. 10. The solid curve provides the EIG values generated by the two-response-category model across a range of *ex/id* decision thresholds. The peak of the curve is at the same position as the optimal threshold (green vertical line) from the top panel of Fig. 3. The blue, dotted horizontal line shows the model EIG, that is, using the decision thresholds that best account for examiner data, from the three-response-category model. The green, dashed horizontal line shows the optimal EIG from the three-response-category model. In short, regardless of the location of the decision threshold, three-response categories can produce substantially more information about the whether the prints are from the same source than two-response categories.

It is important to note that, although this work demonstrates a clear advantage of the five-response-category scale, it is relatively agnostic to the terms used to label those categories (e.g., source identification vs extremely strong support for common source; OSAC 2018; ASAB 2024). The choice of the decision labels has an impact both on how the examiners behave and how they are interpreted by the consumer (Busey et al. 2022). Thus, the examiner data analyzed here, which was collected with concrete decision labels, reflect the use of those labels. The simulations exploring optimal responding, however, are not influenced by label choice. That is, the SDT model thresholds were determined solely based on optimization of EIG. Regardless, care should be taken when deciding on labels for categories. In particular, additional work is needed to determine which decision labels induce the most effective use of the response scale, that is, which decision labels encourage responding that is well calibrated against the actual strength of the evidence. That said, the five-response-category labels used here (exclusion, support for different source, inconclusive, support for same source, identification), do produce close-to-optimal responding, and so are likely a good candidate as a starting point. However, several authors have made persuasive arguments against definitive conclusions (Champod et al. 2016;

[11] This analysis assumes that the distributions are fixed across response category conditions. Median parameter values from the Bayesian SDT model are used throughout.

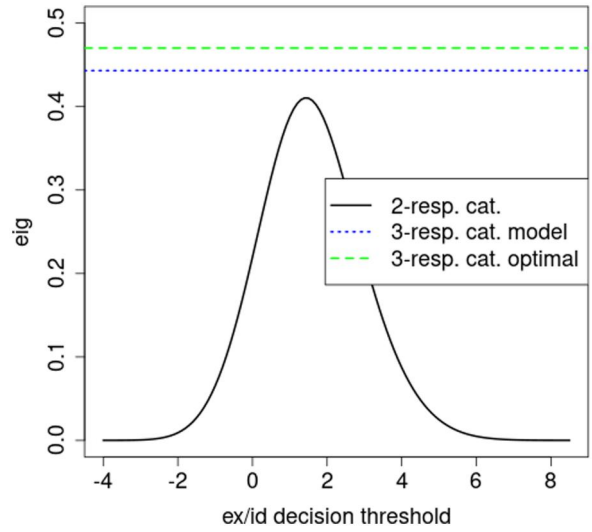

Figure 10. Comparison of the EIG values generated by the two-response-category condition for different *ex/id* decision thresholds and the EIG values generated by the three-response-category condition with fixed model and optimal decision thresholds.

Biedermann 2022). In addition, the terms should be calibrated against the actual strength of the evidence, both in terms of the intention of the examiner and the understanding by a consumer.

As a first step toward addressing the effect of label choice on EIG, we compare the response data of fingerprint examiners who used different sets of five category responses on the same prints (Busey et al. 2022). The first set of response categories was the expanded traditional scale used in the previous analyses (exclusion, support for different sources, inconclusive, support for same source, identification). The second set was the strength-of-support scale (extremely strong support for different sources, support for different sources, inconclusive, support for common source, extremely strong support for common source). For these scales, the EIG values were very similar, 0.479 and 0.490 for the expanded traditional and strength-of-support scales, suggesting robust results, although additional work is required to address this issue across different types of scales.

Although the SDT model used here is widely used and does an excellent job reproducing the empirical results, it does make a number of assumptions that deserve to be made explicit and addressed in future work. First, the evidence distributions were assumed to be normally distributed (Wixted 2020). Second, the current SDT model assumes that each fingerprint pair produces a single evidence value, for example, a level of support for the same source proposition. It may be that fingerprints are compared on multiple dimensions (Rotello, Macmillan, and Reeder 2004), for example, print pairs may be assessed on both appropriateness for comparison and support for same source, and these dimensions may interact necessitating a more complex model. For example, saying "of value for identification" less often would produce overall greater discriminability for the remaining examined pairs. Finally, all decision variability was assumed to be captured by the evidence distributions. A more complete model would explicitly incorporate separate effects of individuals and labs (Luby, Mazumder, and Junker 2020; Luby 2023).

Note that we also modeled the examination process at the level of the individual examiner. A laboratory report is the conglomerate of multiple examination decisions, not produced by a single examiner, but rather by a team of examiners acting according to an examination protocol and subject to verification. Thus, this work addresses a single stage of a larger process. For example, the placement of an individual examiner's threshold can vary based on their psychological state, colleagues, and the expectations and policies of their laboratories. Furthermore, identification decisions undergo an additional step before reporting, known as verification, where a colleague performs a comparison and decision-making process as well. It is the combination of these two examiners' opinions that results in the report to the judicial system. Because examiners work in close proximity for many years, they gain a sense of the thresholds of their

colleagues and may adjust their own internal thresholds in response. The final decision is a product of this more complex process and future work should explore these dependencies.

The present analyses did not consider the costs of various errors, except indirectly by acknowledging the outsized importance of the identification threshold in our discussion of the analyses. This issue is complicated, because examiners typically do not know how their decisions will be used in the judicial system, which makes it difficult to anticipate the costs of errors. Developing approaches to navigate these challenges is an important goal for future research, and information-based measures will likely play a critical role in these developments. Evidence-gathering procedures only lead to costs or benefits to the extent that they convince triers-of-fact to take different actions than they would have otherwise. For example, an incorrect fingerprint identification does not directly put an innocent person in jail, but it might help to convince a juror that an innocent person is guilty. EIG is a theoretically driven measure for representing both the cost of misleading evidence and the value of valid evidence, and therefore represents the fundamental properties that determine the usefulness of an evidence-gathering procedure: helping people who need to make important decisions become more certain of what is actually true. Given that examiner reports with high information value are likely to be useful under a wide range of cost scenarios, the information-based approach provides a robust guide to effective evidence-gathering procedures, such as fingerprint examinations.

8. Conclusion

Using EIG within a SDT modeling framework, the results clearly advocate for the use of a five-response-category decision scale. The five-response-category scale produces an overall higher information gain, maintains high information gain even for risk-averse examiners, is more tolerant of non-optimal decision thresholds, and tends to encourage examiner decision thresholds that are very close to optimal. More generally, our work illustrates the critical role that information measures can play when evaluating forensic data.

Conflict of interest statement. None declared.

Data availability

All code and results are provided at https://osf.io/ugj2q or https://doi.org/10.17605/OSF.IO/UGJ2Q.

Appendix A: EIG Details

In this appendix, we provide a brief introduction to the mathematics underlying the calculation of EIG (see Starns, Cohen, and Rotello 2023, for more information).

A.1 Data

We work through an example using the three-response-category data examined in the main text (Busey et al. 2022). These data are provided in Table A1. The shaded region of Table A1 provides the joint probabilities for the experimental data. For example, 32.6% of all trials were non-mated pairs that resulted in an exclusion decision.

Table A1. Decision proportions for the three-response-category data from Busey et al. (2022).

	Decision			
Pair type	Exclusion	Inconclusive	Identification	Total
Non-mated	0.326	0.160	0.004	**0.489**
Mated	0.035	0.198	0.278	**0.511**
Total	**0.360**	**0.358**	**0.282**	**1.000**
***p*(Non-mated | Resp.)**	**0.904**	**0.447**	**0.013**	
***p*(Mated | Resp.)**	**0.096**	**0.553**	**0.987**	

Note that 50% of the prints presented for examination were mated and 50% were non-mated. Examiners were given the option to label a pair "not of value for examination". The data in Table A1 are restricted to the "of-value" responses.

A.2 Bayesian updating

Bayes rule was used to compute the posterior probability over mated and non-mated print pairs. For example, consider the probability that an excluded pair was non-mated. Because the data are joint probabilities, this probability can be computed as

$$p(n|ex) = \frac{p(n \cap ex)}{p(ex)} = \frac{0.326}{0.360} = 0.904, \tag{A1a}$$

where n is non-mated, m is mated, and ex is an exclusion decision. In general, using Bayes rule,

$$p(n|ex) = \frac{p(ex|n)p(n)}{p(ex|n)p(n) + p(ex|m)p(m)} = \frac{0.666 \times 0.489}{0.666 \times 0.489 + 0.068 \times 0.511} = 0.904. \tag{A1b}$$

A.3 Surprisal

Let Y be a discrete random variable with outcomes y_j, each with probability $p(y_j)$. In the current context, Y has two possible outcomes: non-mated, n, associated with $p(n)$, and mated, m, associated with $p(m)$.

Shannon (1948) derived the equation for the amount of information gained by revealing a single outcome, that is, finding out about the true state of the world. This measure is based on a value called *surprisal*, which is typically defined as

$$I(y_j) = \log_2\left(\frac{1}{p(y_j)}\right). \tag{A2a}$$

When using log base 2, surprisal is measured in *bits*. Intuitively, surprisal measures how "surprised" you would be at learning about the true state of the world, with higher values indicating more surprise. Judgments that assign a lower probability to the true state of the world have higher surprisal values. As such, outcomes that tend to decrease surprisal produce positive information gain values; that is, they tend to more closely align judgments (e.g., the assessed probability that a given pair of latent and comparison prints is mated or non-mated) with the true state of the world.

For example, consider a single fingerprint-pair comparison that comes from the pool of comparisons that produced the data in Table A1. Assume that this comparison involved a non-mated pair but the examiner returned an identification decision. Someone who knows the examiner response but not the true relationship between the latent and comparison print would reasonably be 98.7% certain that the pair is mated (0.278/0.282 of identifications were for mated pairs), leaving only a 1.3% chance that the pair is non-mated. In this case, however, the examiner response was misleading, and someone with access to the examiner report would be very surprised to learn that this particular pair is actually non-mated, as indicated by a high surprisal value

$$I(id|n) = \log_2\left(\frac{1}{0.013}\right) = 6.265. \tag{A2b}$$

In contrast, consider another fingerprint-pair comparison that comes from the data in Table A1. Assume that this comparison was also for a non-mated pair, but this time the examiner returned an exclusion decision. Someone who knows the examiner response but not the true relationship between the latent and comparison print would reasonably be 90.4% certain that the pair is **non**-mated (.326/.360 exclusions were for non-mated pairs). The examiner

response was a reliable guide in this example, and someone with access to the examiner report would not be very surprised to learn that this particular pair was non-mated, indicated by a small surprisal value

$$I(ex|n) = \log_2\left(\frac{1}{0.904}\right) = 0.146. \tag{A2c}$$

Informative examiners tend to provide reliable information; that is, they tend to help people with access to the examiner report assign a high probability to the true state of the world, as indexed by low surprisal values.

A.4 Entropy

Entropy, H, is the average surprisal across a probability distribution. That is, entropy is the summed surprisal over all possible outcomes, where the surprisal of each outcome is weighted by the probability of that outcome

$$H(Y) = \sum\nolimits_j p(y_j)\log_2\left(\frac{1}{p(y_j)}\right). \tag{A3a}$$

Outcomes with probability zero are dropped.

Entropy measures the uncertainty associated with a probability distribution—the less predictable the outcome, the higher the entropy. Entropy embodies the intuition that it is easier to judge whether a pair of prints are mated or not when those prints are sampled from a set of predominantly mated or predominantly non-mated pairs. It is hardest to judge whether a pair of prints are mated when those prints are sampled from an even mix of mated and non-mated pairs.

Continuing with examples from Table 1A, entropy would be defined not for a single fingerprint comparison, as in the surprisal examples, but for an entire set of comparisons. For example, entropy can be defined for all of the comparisons in aggregate. Given that the examiners provided judgments for 50% mated and 50% non-mated pairs, the optimal strategy, before seeing the examiner's decision, is to judge that every print has a 50% chance of being mated or non-mated.[12] The entropy is then given by

$$\begin{aligned} H(Y) &= p(n)\log_2\left(\frac{1}{p(n)}\right) + p(m)\log_2\left(\frac{1}{p(m)}\right) \\ &= 0.50 \times \log_2\left(\frac{1}{0.50}\right) + 0.50 \times \log_2\left(\frac{1}{0.50}\right) \\ &= 0.50 \times 1.00 + 0.50 \times 1.00 = 1.00 \ \textit{bit}. \end{aligned} \tag{A3b}$$

The value of 1 represents the 1 bit of information that would be gained, on average, if it was revealed whether the prints are mated or not. Because this value was determined before knowledge of the examiner decision, it is called the prior entropy.

Entropy can also be defined for the subset of comparisons that are associated with a given examiner decision, and this value represents the posterior entropy that takes the examiner decision into account. The posterior entropy is determined in the same way as the priori entropy, but uses the posterior probabilities of a mated and non-mated pair as given in Table A1 and discussed in the Bayesian updating section. For example, say an examiner provided an identification decision (*id*) to a pair of prints. Then the posterior entropy is given by

[12] Recall that the data in Table A1 deviate slightly from a 50/50 base rate due to the selection of only "of-value" trials.

$$\begin{aligned} H(Y|id) &= p(n|id)\log_2\left(\frac{1}{p(n|id)}\right) + p(m|id)\log_2\left(\frac{1}{p(m|id)}\right) \\ &= 0.013 \times \log_2\left(\frac{1}{0.013}\right) + 0.987 \times \log_2\left(\frac{1}{0.987}\right) \\ &= \ 0.013 \times 6.265 + 0.987 \times 0.019 = 0.098 \ \ bit. \end{aligned} \tag{A3c}$$

Now, only 0.098 bits of information are gained on average from learning whether the print pair is mated after an identification decision. This change occurs because identifications almost always stem from mated pairs (98.7%), and very little information is gained if it is learned that the prints are mated. More information is gained if it is learned the prints are non-mated, but this happens very seldom.

A.5 Information Gain

The information gain associated with a particular examiner decision is the difference in entropy before and after knowledge of the examiner decision. That is, the information gain is the difference between the prior entropy and posterior entropy. Continuing the example from the previous section, the information gain associated with an identification response is given by

$$IG(id) = \ H(Y)\text{-}H(Y|id) = 1.000\text{-}0.098 = 0.901. \tag{A4}$$

That is, 0.901 bits of information are gained from an identification decision, on average.

A.6 EIG

Information gain is computed for a single response. EIG takes the weighted average of information gain across all responses. EIG is computed as

$$EIG = \sum\nolimits_i p(d_i) IG(d_i), \tag{A5}$$

where d_i is Decision i. For the data from Table A1, EIG is 0.453. The complete steps for computing EIG for the data from Table A1 are provided in Table A2.

Table A2. Steps in computing EIG for the three-decision category data from Table A1.

Decision (d)	$p(d\|n)$	$p(d\|m)$	$p(d)$	$p(m\|d)$	$H(Y\|d)$	$IG(d)$	$p(d)\times$IG(d)
Exclusion	0.666	0.068	0.360	0.096	0.456	0.543	0.196
Inconclusive	0.327	0.388	0.358	0.553	0.992	0.008	0.003
Identification	0.007	0.544	0.282	0.987	0.098	0.901	0.254
						$\sum p(d) \times IG(d) = EIG = 0.453$	

Notes. $p(d|n)$ is probability of Decision d on comparisons of non-mated prints; $p(d|m)$ is probability of Decision d on comparisons of mated prints; $p(d)$ is overall probability of Decision d; p$(m|d)$ is posterior probability of a mated pair for Decision d (posterior probability of non-mated is $1 - p(m|d)$); $H(Y|d)$ is entropy given Decision d; $IG(d)$ information gain for Decision d; EIG is expected information gain. These calculations assume prior proportions of 0.50 for both non-mated and mated prints.

Appendix B: Analysis details

This appendix provides details of the Bayesian implementation of the signal detection model of fingerprint analysis discussed in the text. All code and results are provided at https://osf.io/ugj2q or https://doi.org/10.17605/OSF.IO/UGJ2Q.

B.1 Software

The model was developed and run using the rstan package (Stan Development Team 2023 v2.21.8) in R (R Core Team 2023; v4.3.2). HDIs were computed using the package HDInterval package (Meredith and Kruschke 2022, v0.2.4).

B.2 MCMC parameters

Four chains of 10,000 iterations were run, including 5,000 warm-ups.

B.3 Data

The model was simultaneously applied to the three- and five-response-category data from Busey et al. (2022). Although they provided a decision for all fingerprint pairs, examiners initially determined whether each pair was of good enough quality or "of-value" for comparison. Only responses initially deemed "of-value" for exclusion or identification were used. The same parameters were also used to account for a two-response-category condition, which was generated from the three-response-category condition by combining the exclusion and inconclusive categories into a single non-id response category.

B.4 Prior distributions

For both the three- and five-response-category conditions, we assumed, without loss of generality, that:

$$\mu_{nm} = 0.00, \text{ mean of the non-mated distribution.}$$
$$\sigma_{nm} = 1.00, \text{ sd of the non-mated distribution.}$$

The following are the prior distributions for the mated distribution parameters. We assumed these parameters are shared across the three- and five-response-category conditions:

$$\mu_m \sim N(2.40,\ 4.00), \text{ mean of the mated distribution.}$$
$$\sigma_m \sim gamma(1.00,\ 1.00), \text{ sd of the mated distribution.}$$

Minimal allowed values for μ_m and σ_m were 0.10.

The following are the prior distributions for the decision thresholds. This threshold parameterization was selected to reduce parameter dependencies. These thresholds were transformed into more meaningful thresholds as discussed below.

$c3_{bias} \sim N(0.00,\ 3.00)$, center of the inconclusive category relative to the cross-over point of the non-mated and mated distributions, for the three-response-category condition. For example, if $c_{bias3} = 0$, the inconclusive category is centered on the cross-over point of the non-mated and mated distributions.

$c3_{inc} \sim U(0.01,\ 6.00)$, width of the inconclusive response category, for the three-response-category condition.

$c5_{bias} \sim N(0.00,\ 3.00)$, center of the inconclusive category relative to the cross-over point of the non-mated and mated distributions, for the five-response-category condition.

$c5_{inc} \sim U(0.01,\ 6.00)$, width of the inconclusive response category, for the five-response-category condition.

$c5_{sds} \sim U(0.01,\ 6.00)$, width of the support for different source response category, for the five-response-category condition.

$c5_{sss} \sim U(0.01,\ 6.00)$, width of the support for same source response category, for the five-response-category condition.

Minimal allowed values for the *bias* decision thresholds were 0.10, and minimal values for all other decision thresholds were 0.01.

B.5 Decision thresholds transformations

The parameters from the previous section were transformed to the decision thresholds discussed in the main text as follows. *ex* is the exclusion response category, *sds* is support for different source, *inc* is inconclusive, *sss* is support for same source, *id* is identification, and the *non-id* category was created by combining the *ex* and *inc* response categories from the three-response-category condition. For the three-response-category condition:

$$c3_{ex/inc} = c3_{bias} - {}^{c3_{inc}}/_{2},$$
$$c3_{inc/id} = c3_{bias} + {}^{c3_{inc}}/_{2}.$$

For the five-response-category condition:

$$c5_{ex/sds} = c5_{bias} - {}^{c5_{inc}}/_{2} - c5_{sds},$$
$$c5_{sds/inc} = c5_{bias} - {}^{c5_{inc}}/_{2},$$
$$c5_{inc/sss} = c5_{bias} + {}^{c5_{inc}}/_{2},$$
$$c5_{sss/id} = c5_{bias} + {}^{c5_{inc}}/_{2} + c5_{sss}.$$

For the two-response-category condition there is only one decision threshold, which is identical to the three-response-category *inc/id* decision threshold:

$$c2_{non\text{-}id/id} = c3_{inc/id}.$$

B.6 Response proportions

Response proportions were computed as the area under the appropriate distribution between the relevant decision thresholds. In the following, $\Phi(X;\mu,\sigma)$ is the cumulative Normal distribution to X with mean μ and sd σ, *ex* is the exclusion response category, *sds* is support for different source, *inc* is inconclusive, *sss* is support for same source, *id* is identification, and *non-id* was created by combining the ex and *inc* response categories from the three-response-category condition. For the two-response-category condition:

$$p(non\text{-}id|non\text{-}mated) = \Phi(c2_{non\text{-}id/id};\mu_{nm},\sigma_{nm})$$
$$p(id|non\text{-}mated) = 1\text{-}\Phi(c2_{non\text{-}id/id};\mu_{nm},\sigma_{nm})$$
$$p(non\text{-}id|mated) = \Phi(c2_{non\text{-}id/id};\mu_{m},\sigma_{m})$$
$$p(id|mated) = 1\text{-}\Phi(c2_{non\text{-}id/id};\mu_{m},\sigma_{m})$$

For the three-response-category condition:

$$p(ex|non\text{-}mated) = \Phi(c3_{ex/inc};\mu_{nm},\sigma_{nm})$$
$$p(inc|non\text{-}mated) = \Phi(c3_{inc/id};\mu_{nm},\sigma_{nm})\text{-}\Phi(c3_{ex/inc};\mu_{nm},\sigma_{nm})$$
$$p(id|non\text{-}mated) = 1\text{-}\Phi(c3_{inc/id};\mu_{nm},\sigma_{nm})$$
$$p(ex|mated) = \Phi(c3_{ex/inc};\mu_{m},\sigma_{m})$$
$$p(inc|mated) = \Phi(c3_{inc/id};\mu_{m},\sigma_{m})\text{-}\Phi(c3_{ex/inc};\mu_{m},\sigma_{m})$$
$$p(id|mated) = 1\text{-}\Phi(c3_{inc/id};\mu_{m},\sigma_{m})$$

For the five-response-category condition:

$$p(ex|non\text{-}mated) = \Phi(c5_{ex/sds}; \mu_{nm}, \sigma_{nm})$$

$$p(sds|non\text{-}mated) = \Phi(c5_{sds/inc}; \mu_{nm}, \sigma_{nm})\text{-}\Phi(c5_{ex/sds}; \mu_{nm}, \sigma_{nm})$$

$$p(inc|non\text{-}mated) = \Phi(c5_{inc/sss}; \mu_{nm}, \sigma_{nm})\text{-}\Phi(c5_{sds/inc}; \mu_{nm}, \sigma_{nm})$$

$$p(sss|non\text{-}mated) = \Phi(c5_{sss/id}; \mu_{nm}, \sigma_{nm})\text{-}\Phi(c5_{inc/sss}; \mu_{nm}, \sigma_{nm})$$

$$p(id|non\text{-}mated) = 1\text{-}\Phi(c5_{sss/id}; \mu_{nm}, \sigma_{nm})$$

$$p(ex|mated) = \Phi(c5_{ex/sds}; \mu_{m}, \sigma_{m})$$

$$p(sds|mated) = \Phi(c5_{sds/inc}; \mu_{m}, \sigma_{m})\text{-}\Phi(c5_{ex/sds}; \mu_{m}, \sigma_{m})$$

$$p(inc|mated) = \Phi(c5_{inc/sss}; \mu_{m}, \sigma_{m})\text{-}\Phi(c5_{sds/inc}; \mu_{m}, \sigma_{m})$$

$$p(sss|mated) = \Phi(c5_{sss/id}; \mu_{m}, \sigma_{m})\text{-}\Phi(c5_{inc/sss}; \mu_{m}, \sigma_{m})$$

$$p(id|mated) = 1\text{-}\Phi(c5_{sss/id}; \mu_{m}, \sigma_{m})$$

B.7 Computation of optimal and maximum EIG

For a given μ_m and σ_m, the *optimal* decision thresholds were determined by maximizing EIG using the optim optimization function in R.

For a given μ_m and σ_m, the *maximum* EIG was estimated by drawing 10,000 samples from the non-mated and mated distributions, computing the likelihood ratio that each sample was drawn from the mated distribution, transforming these likelihood ratios to the probability of a mated pair, and then computing EIG over the probabilities as described previously.

See code for further details.

B.8 Diagnostics and posterior predictive check

All diagnostics were checked via inspection of trace plots. All parameters converged. Pairwise posterior samples for the model parameters (including transformed decision thresholds) are shown in Fig. B1.

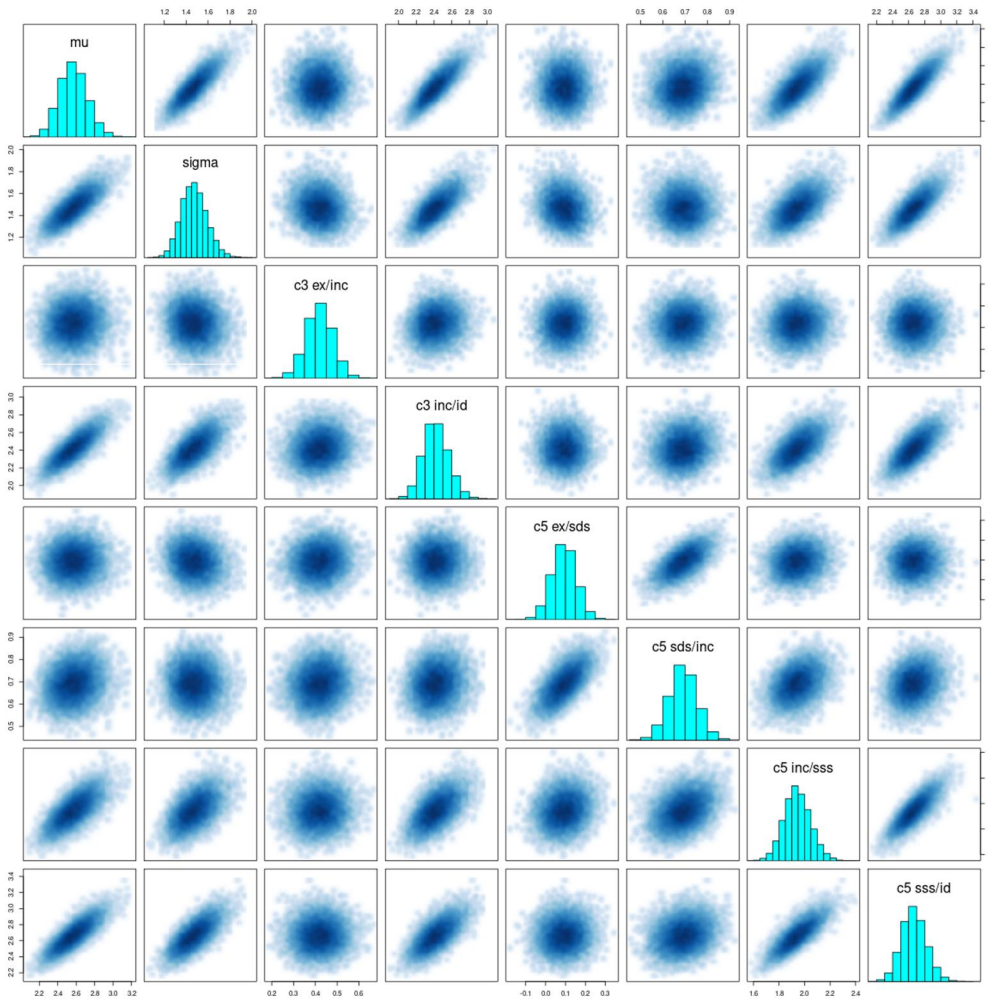

Figure B1. Pairwise posterior samples from the Bayesian model. *mu* and *sigma* are the mean and standard deviation of the mated distribution, respectively. *c3* and *c5* indicate the three- and five-response-category conditions, respectively. *ex* is the exclusion response category; *sds* is support for different source; *inc* is inconclusive; *sss* is support for same source; and *id* is identification. For example, *c3 ex/inc* is the decision threshold separating the exclusion and inconclusive response categories for the three-response-category condition.

A posterior predictive check shows that the model also did a very good job accounting for the data from all response category conditions, as illustrated in Fig. B2.

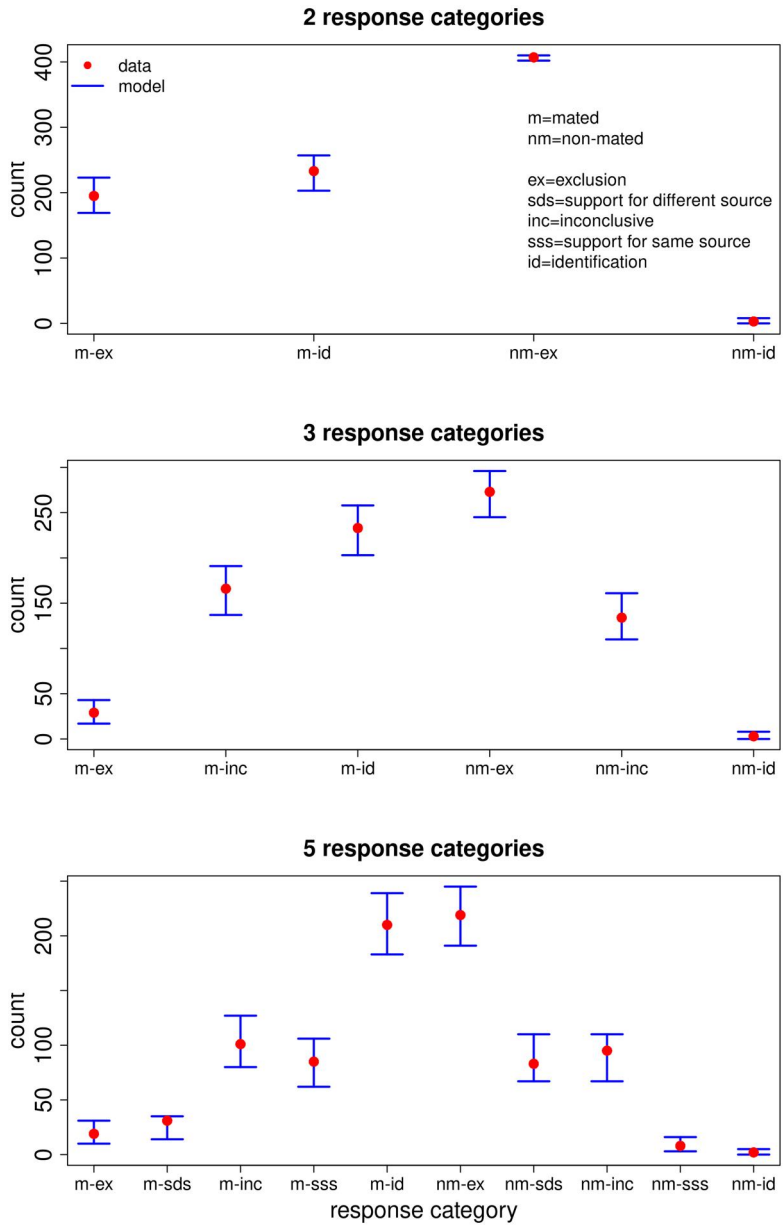

Figure B2. Posterior predictive results from the Bayesian implementation of the signal-detection model for the two-, three-, and five-response-category conditions. The model results are 95% HDIs.

References

Albright, T. D. (2022). 'How to Make Better Forensic Decisions', *Proceedings of the National Academy of Sciences*, **119**: e2206567119.

ANSI/ANAB. (2024). 'Best Practice Recommendation for Comparison and Evaluation of Friction Ridge Impressions', https://www.aafs.org/sites/default/files/media/documents/166_BPR_e1.pdf

Banks, W. P. (1970). 'Signal Detection Theory and Human Memory', *Psychological Bulletin*, **74**: 81–99. https://doi.org/10.1037/h0029531
Benish, W. A. (1999). 'Relative Entropy as a Measure of Diagnostic Information', *Medical Decision Making*, **19**: 202–6. https://doi.org/10.1177/0272989x9901900211
Benjamin, A. S., Tullis, J. G., and Lee, J. H. (2013). 'Criterion Noise in Ratings-Based Recognition: Evidence from the Effects of Response Scale Length on Recognition Accuracy', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **39**: 1601–1608. https://doi.org/10.1037/a0031849
Biedermann, A. (2022). 'The Strange Persistence of (Source) "Identification" Claims in Forensic Literature through Descriptivism, Diagnosticism and Machinism', *Forensic Science International: Synergy*, **4**: 100222. https://doi.org/10.1016/j.fsisyn.2022.100222
Busey, T., and Coon, M. (2023). 'Not all Identification Conclusions Are Equal: Quantifying the Strength of Fingerprint Decisions', *Forensic Science International*, **343**: 111543.
Busey, T. et al. (2022). 'Validating Strength-Of-Support Conclusion Scales for Fingerprint, Footwear, and Toolmark Impressions', *Journal of Forensic Sciences*, **67**: 936–54.
Campbell, W. M. et al. (2005). 'Estimating and evaluating confidence for forensic speaker recognition', Proceedings (ICASSP '05). *IEEE International Conference on Acoustics, Speech, and Signal Processing*. https://doi.org/10.1109/icassp.2005.1415214aq_pubdet
Carter, K. E. et al. (2020). 'The Utility of Expanded Conclusion Scales during Latent Print Examinations', *Journal of Forensic Sciences*, **65**: 1141–54.
Chaitin, G. J. (1975). 'A theory of Program Size Formally Identical to Information Theory', *Journal of the ACM*, **22**: 329–340. https://doi.org/10.1145/321892.321894
Champod, C. et al. (2016). *Fingerprints and Other Ridge Skin Impressions* (2nd ed.). CRC Press.
Cover, T. M., and Thomas, J. A. (2006). *Elements of Information Theory*. J. Wiley.aq_publoc
Gardner, B. O., Neuman, M., and Kelley, S. (2021). 'Latent Print Quality in Blind Proficiency Testing: Using Quality Metrics to Examine Laboratory Performance', *Forensic Science International*, **324**: 110823. https://doi.org/10.1016/j.forsciint.2021.110823
Garner, W. R. (1953). 'An Informational Analysis of Absolute Judgments of Loudness', *Journal of Experimental Psychology*, **46**: 373–80. https://doi.org/10.1037/h0063212
Gong, D. et al. (2015). 'A maximum entropy feature descriptor for age invariant face recognition', *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2015.7299166
Haase, S.J., Theios, J. and Jenison, R. (1999). 'A Signal Detection Theory Analysis of an Unconscious Perception Effect', *Perception & Psychophysics*, **61**: 986–92. https://doi.org/10.3758/BF03206912
Luby A. (2023). 'A Method for Quantifying Individual Decision Thresholds of Latent Print Examiners', *Forensic Science International: Synergy*, **7**: 100340. https://doi.org/10.1016/j.fsisyn.2023.100340.
Luby, A., Mazumder, A. and Junker, B. (2020). 'Psychometric Analysis of Forensic Examiner Behavior', *Behaviormetrika* **47**: 355–384. https://doi.org/10.1007/s41237-020-00116-6.
Mannering, W. M. et al. (2021). 'Are Forensic Scientists Too Risk Averse?' *Journal of Forensic Sciences*, **66**: 1377–400.
Martire, K. A., Kemp, R. I., and Newell, B. R. (2013). 'The Psychology of Interpreting Expert Evaluative Opinions', *Australian Journal of Forensic Sciences*, **45**: 305–314. https://doi.org/10.1080/00450618.2013.784361.
Meredith. M., and Kruschke, J. (2022). _HDInterval: Highest (Posterior) Density Intervals_. R package version 0.2.4, <https://CRAN.R-project.org/package=HDInterval>.
Mickes, L. (2015). 'Receiver Operating Characteristic Analysis and Confidence–Accuracy Characteristic Analysis in Investigations of System Variables and Estimator Variables That Affect Eyewitness Memory', *Journal of Applied Research in Memory and Cognition*, **4**: 93–102.
Mickes, L., Flowe, H. D., and Wixted, J. T. (2012). 'Receiver Operating Characteristic Analysis of Eyewitness Memory: Comparing the Diagnostic Accuracy of Simultaneous versus Sequential Lineups', *Journal of Experimental Psychology: Applied*, **18**: 361.
Neumann, C., Evett, I. W., and Skerrett, J. (2012). 'Quantifying the Weight of Evidence from a Forensic Fingerprint Comparison: A New Paradigm', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **175**: 371–415. https://doi.org/10.1111/j.1467-985X.2011.01027.
Osterburg, J. W. et al. (1977). 'Development of a Mathematical Formula for the Calculation of Fingerprint Probabilities based on Individual Characteristics', *Journal of the American Statistical Association*, **72**: 772. https://doi.org/10.2307/2286458
OSAC. (2018). 'Standard for Friction Ridge Examination Conclusions', https://www.nist.gov/system/files/documents/2020/03/23/OSAC%20FRS%20CONCLUSIONS%20Document%20Template%202020_Final.pdf.
Phillips, V., Saks, M., and Peterson, J. (2001). 'The Application of Signal Detection Theory to Decision-Making in Forensic Science', *Journal of Forensic Sciences*, **46**: 294t–308. https://doi.org/10.1520/jfs14962j.

Ramos, D., and Gonzalez-Rodriguez, J. (2008). 'Cross-entropy analysis of the information in forensic speaker recognition', *Proceedings of IEEE Odyssey*.aq_pubdet

Ramos, D. et al. (2007). 'Information-theoretical comparison of likelihood ratio methods of forensic evidence evaluation', *Third International Symposium on Information Assurance and Security*. https://doi.org/10.1109/ias.2007.63.aq_pubdet

Rotello, C. M., Macmillan, N. A., and Reeder, J. A. (2004). 'Sum-Difference Theory of Remembering and Knowing: A Two-Dimensional Signal-Detection Model', *Psychological Review*, **111**: 588.

R Core Team (2023). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Shannon, C. E. (1948). 'A Mathematical Theory of Communication', *The Bell system technical journal*, **27**: 379–423.

Smith, A. M., and Neal, T. M. S. (2021). 'The Distinction between Discriminability and Reliability in Forensic Science', *Science & Justice*, **61**: 319–31. https://doi.org/10.1016/j.scijus.2021.04.002.

Stan Development Team (2023). RStan: the R interface to Stan. R package version 2.21.8. https://mc-stan.org/.

Starns, J. J., Cohen, A. L., and Rotello, C. M. (2023). 'A Complete Method for Assessing the Effectiveness of Eyewitness Identification Procedures: Expected Information Gain', *Psychological Review*, **130**: 677–719. https://doi.org/10.1037/rev0000332.

Swofford, H. J. et al. (2018). 'A Method for the Statistical Interpretation of Friction Ridge Skin Impression Evidence: Method Development and Validation', *Forensic Science International*, **287**: 113–26. https://doi.org/10.1016/j.forsciint.2018.03.043.

Thompson, W. C. (2023). 'Shifting Decision Thresholds can Undermine the Probative Value and Legal Utility of Forensic Pattern-Matching Evidence', *Proceedings of the National Academy of Sciences*, **120**. https://doi.org/10.1073/pnas.2301844120

Thurstone, L. L. (1927). 'A Law of Comparative Judgment', *Psychological Review*, **34**: 273.

Wixted, J. T. (2020). 'The Forgotten History of Signal Detection Theory', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **46**: 201–233. https://doi.org/10.1037/xlm0000732.

Wixted, J. T. and Mickes, L. (2018). 'Theoretical vs. Empirical Discriminability: The Application of ROC Methods to Eyewitness Identification', *Cognitive Research: Principles and Implications*, **3**: 1–22.

Wixted, J. T. and Wells, G. L. (2017). 'The Relationship between Eyewitness Confidence and Identification Accuracy: A New Synthesis', *Psychological Science in the Public Interest*, **18**: 10–65.

Zane, A. et al. (2025). 'Bayesian Sequential Experimental Design for Planning Series of Police Lineups', *Law, Probability, and Risk*, **24**. Advance online publication. https://doi.org/10.1093/lpr/mgae017.

Law, Probability and Risk, 2025, 24, 1–30
https://doi.org/10.1093/lpr/mgaf004
Research Article