



Tracking the growth of visual evidence in fingerprint comparison tasks

Thomas Busey¹ · Brandi Emerick¹ · John Vanderkolk²

Accepted: 8 October 2022 / Published online: 30 November 2022
© The Psychonomic Society, Inc. 2022

Abstract

Fingerprint comparisons are extended in time due to the fine details (minutiae) that necessitate multiple eye fixations throughout the comparison. How is evidence accumulated across these multiple regions? The present work measures decisions at multiple points during a comparison to address how feature diagnosticity and image clarity play a role in evidence accumulation. We find that evidence is accumulated at a constant rate over time, with evidence for identification and exclusion accumulated at similar rates. Manipulations of image diagnosticity and image clarity demonstrate two exceptions to this constant rate: Highly diagnostic evidence followed by weak evidence tends to lose the initial benefits of the strong start, and low image clarity at the start of the comparison can be overcome with high image clarity at the end of the comparison. The results suggest that examiners tend to treat each region fairly independently (as demonstrated by linear evidence accumulation), with only weak evidence for hysteresis effects that tend to fade as additional regions are presented. Data from transition probability matrices support an incremental evidence accumulation account, with very little evidence for rapid “aha” moments even for exclusion decisions. The results are consistent with a model in which each fixated region contributes an independent unit of evidence, and these accumulate to form an eventual decision. Fingerprint comparisons do not seem to depend on which regions are selected first, and thus examiners need not worry about finding the most diagnostic region first, but instead focus on conducting a complete analysis of the latent print.

Keywords Evidence accumulation · Signal detection theory · Expertise · Saliency · Forensics · Fingerprints

Introduction

The majority of fingerprint examinations are not done by computer, but instead by human experts. There are no mandated criteria for sufficiency for their decisions (SWGFAST, 2013), and laypersons tend to over-interpret these decisions (Swofford & Cino, 2017). Thus, it is important to understand how examiners accumulate information during these tasks, which has a direct bearing on their ultimate decision.

Fingerprint examiners receive latent fingerprints collected from crime scenes and compare these against exemplar

fingerprints from known sources. These exemplars come either from suspects or are returned from database searches. The examiner conducts an analysis of the latent impression and then performs a comparison with one or more exemplars using a process known as ACE-V (SWGFAST, 2013; Tierney, 2013). During the analysis phase, an examiner identifies individual regions or features for later comparison. During the comparison phase, the examiner assesses the amount of perceived detail in agreement for areas they determine might correspond. Finally, this evidence is accumulated and evaluated against an external standard to reach one of three conclusions: In an Exclusion conclusion, the examiner expresses their expert opinion that the two impressions originated from different sources; in an Identification conclusion, the examiner expresses their opinion that the two impressions come from the same source; if neither conclusion can be reached, the examiner can give an Inconclusive conclusion (and in some cases ask for better exemplars if exemplar quality is the limiting factor). After a conclusion is made, the comparison is submitted for technical review and in some labs a second examiner repeats some or all of the comparison to serve as verification.

Perhaps surprisingly, there is no fixed standard for what constitutes sufficiency to make an Identification decision in the USA,

John Vanderkolk is a retired member of Indiana State Police Laboratory, Fort Wayne, IN, USA.

✉ Thomas Busey
busey@indiana.edu

¹ Department of Psychological and Brain Sciences, Program in Cognitive Science, Indiana University, 1101 E. 10th St, Bloomington, IN 47405, USA

² Indiana State Police Laboratory, Fort Wayne, IN, USA

nor is there a well-specified set of features that are used (although some efforts have described an extended feature set; see Taylor et al., 2013). Instead, an examiner is free to use whatever information they deem diagnostic and maintain their own internal decision threshold for sufficiency. Although there is no fixed standard in the USA for the number of corresponding features, examiners often describe using 12 or more features before they are completely comfortable making an identification decision (Ulery et al., 2014). Some examiners report relying on a “one unexplainable discrepancy” rule as a basis for an Exclusion, while others describe relying on the totality of the evidence.

Although the present work does not use eye-tracking methods, it is helpful to consider data from eye-gaze recordings to visualize how information may be acquired and integrated. During the comparison process, saccades are made both within and between impressions as illustrated by the sample eye-tracking data in the top panel of Fig. 1 from Busey et al. (2015). One reason for the multiple fixations may be that the relevant details for comparison (minutiae) require foveal inspection. Because multiple regions are visited, the growth of visual information is extended over time, and the fixations that result from each saccade reveal new visual features at a rate that makes it difficult to measure using something like a talk-aloud protocol. In addition, eye gaze measures where the eyes point, not what information the examiner acquires from that location. Thus, neither talk-aloud protocols nor eye tracking provide a complete account of the accumulation of evidence during fingerprint comparisons. While there is evidence that experts can make some basic judgments about fingerprints with presentations as brief as 250 ms (Searston & Tangen, 2017; M. B. Thompson et al., 2014), typical comparisons may take tens of minutes to hours to complete. Whether this is a continuous process or involves an all-or-none “aha” moment has not been addressed in the literature, and the nature and rate of this accumulation is poorly understood.

The goal of the present work is to determine how evidence is accumulated over time during a fingerprint comparison, and to characterize the role of factors such as feature inter-dependency, region diagnosticity, and image clarity in the decision process. Our method provides for investigations of dependencies between sequential decisions, which can be grouped roughly into two categories. The first are perceptually based dependencies, and include effects such as configural processing (Fific & Townsend, 2010; Richler et al., 2015), or failures of perceptual independence or perceptual separation (Ashby & Townsend, 1986). The second are decision-based dependencies, which include processes such as anchoring effects and other decision biases. We discuss each of these below.

Perceptual dependencies

Within the related field of face recognition, the general scientific consensus supports the notion that facial features are processed

inter-dependently. Examples are plentiful, but the classic Thatcher effect (P. Thompson, 1980) illustrates the relational dependency of features; work by Tanaka and Farah (1993) demonstrates dependencies between features when identifying faces; and modeling by Ashby and Townsend (1986) points to inter-dependence at the perceptual processing level. The nature of these configural effects is documented in a wide range of tasks by Richler et al. (2012) using the composite task. The exact nature of these dependencies will depend in part on the model adopted by different authors, but most papers argue for a face-as-template approach or the idea that the interpretation of perceptual information from one region is affected by the presence of a nearby region (for review, see Piepers & Robbins, 2012).

Do such effects exist in fingerprint processing by experts? Although faces and fingerprints share similarities in that they are all composed of similar-looking features that differ primarily in their shape and location, examiners tend not to talk in terms of holistic mechanisms. Instead, the language that is used by experts to describe the comparison process tends to focus on individual features that they describe as “minutiae.” Eye-tracking studies with fingerprint experts (Busey et al., 2013; Busey et al., 2015; Busey et al., 2017; Busey et al., 2021; Hicklin et al., 2019) have revealed that experts typically place one or more features from the latent print (termed a *target group*) into visual working memory, and then make a saccade to the exemplar print to search for a similar region that may be *within tolerance*.¹ If such a region is found, it is described as a *corresponding region*. The search continues by selecting and searching additional target groups to determine possible correspondence or discrepancy.

Although there have been efforts to document the relation between the image features and sufficiency (Ulery et al., 2011, 2012; Ulery et al., 2014), the relation between the decision and the physical stimuli is not well described. Models based on feature rarity are fairly accurate at accounting for the regions visited by examiners as measured by eye tracking, illustrating a role for feature diagnosticity (Busey et al., 2017). In a forensic setting, rare features tend to individualize much better than common features, which may account for the success of these models. In addition, as illustrated by eye-tracking data in Fig. 1 from Busey et al. (2015), the regions selected from an impression are not at random, but tend to be close together and organized, suggesting that relational information may play a role, and experts report the use of techniques such as counting ridges between features (see Hicklin et al., 2019, for eye-tracking examples of such ridge-counting behavior). There is some evidence for holistic or configural processing of fingerprints (Busey & Vanderkolk, 2005), although the classic inversion effect test for holistic processing has produced mixed results (Searston & Tangen, 2017; M. B. Thompson et al., 2014;

¹ This is standard terminology for examiners, although in the psychology literature we would express this in terms of a similarity judgment and a criterion that it is evaluated against.

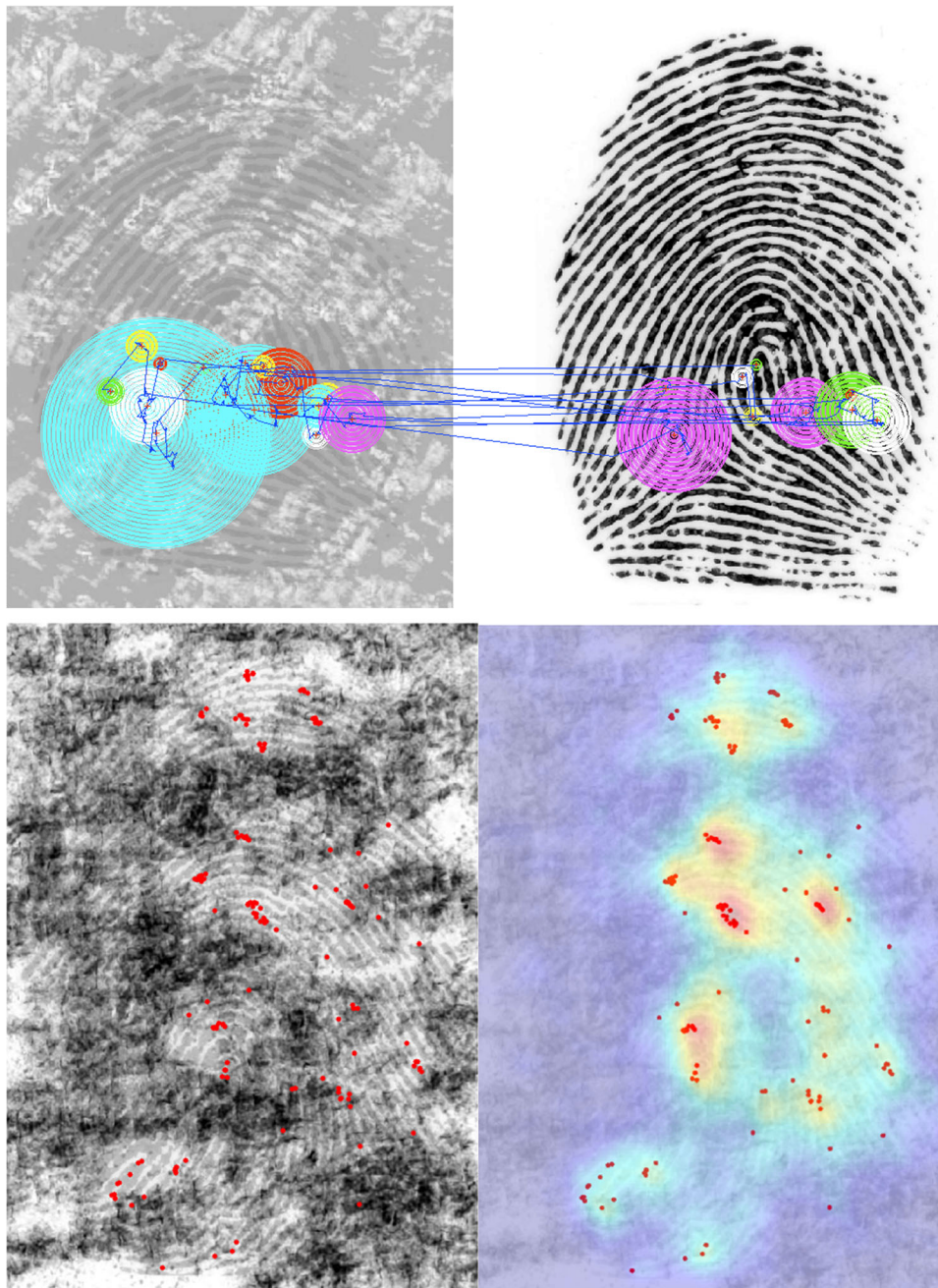


Fig. 1 **Top panel:** Example eye-tracking data from a fingerprint examiner on a comparison with a simulated latent that has been corrupted by multiplicative noise to simulate a latent impression. Circles represent fixations, with circle diameter related to fixation duration. Saccades are represented by blue lines. Data from Busey et al. (2015). **Bottom left**

panel: Minutia marked by 12 latent print examiners during a fingerprint analysis task. **Bottom right panel:** Eye-tracking heatmap overlaid over the marks, illustrating a tight correspondence between looking and marking. Unpublished data from our laboratory

Vogelsang et al., 2017). Thus, the role of relational information is unclear, although anecdotal evidence suggests that experts make use of relational information such as counting ridges or measuring distances to landmarks such as the core or delta regions of an impression. In a closely related target localization task, Hicklin et al. (2019) found that examiners made multiple orienting fixations between a candidate target location and the core area of a fingerprint, presumably to determine the relative

location of the target group. Thus, there may be a deliberative relational process conducted in some cases, which may contribute to feature inter-dependencies.

Decisional dependencies

In addition to perceptual dependencies, an evolving decision process may include dependencies that exist at the decision

stage. Anchoring effects (Mochon & Frederick, 2013; Tversky & Kahneman, 1974) suggest that initial information or decisions create an “anchor” or reference against which subsequent responses are measured against. In perceptual tasks, the sequential effect has long been documented (Holland & Lockhead, 1968), and more recently extended to attractiveness judgments (Kondo et al., 2012). Thus, decision-based dependencies may also exist in a fingerprint comparison task, such that if an examiner initially inspects a region with poor clarity or diagnosticity, this may color the interpretation of subsequent regions, ultimately leading to a different evaluation of the totality of the evidence than if they had initially viewed a high-quality region.

Independent contributions

An alternative to either perceptual or decisional dependencies is that each region adds an additional unit of information, and this information accumulates independently. Such a model would predict a linear relationship between the number of visible regions and a measure of sensitivity such as d' . It would also predict linear increases in the standard deviation of the underlying distribution when modeled by signal detection theory. As more regions become visible, we may encounter diminishing returns in terms of the informativeness of each region, and certainly there will come a point where the entire image is revealed, as in actual casework. However, the selected regions in the present study are relatively small and represent only a small fraction of the available information in a latent fingerprint. Thus, we may not be operating in a region of performance where diminishing returns plays a significant role.

The goal of the present project is to decompose the entire latent print comparison task into a set of individual decisions that will allow us to track the growth of visual information as it accumulates to an ultimate decision. This will allow us to address the following questions:

- How does evidence accumulate for identification and exclusion conclusions, and does evidence accumulate at the same rates for both conclusions?
- How does the diagnosticity or image clarity of the different regions affect the conclusions or evidence accumulation?
- Does information from one region affect the interpretation of another region as might be expected by perceptual grouping models (Kim et al., 2021)?

To answer these questions, we slow down the comparison task and get multiple measures of the accumulated evidence as examiners work toward a conclusion. Our task has two parts: a feature-marking task and a comparison task. In the first part, we emulated the comparison phase by showing examiners

clean impressions and asking them to select eight individual regions in order of feature diagnosticity for purposes of comparison. They repeated this process for 72 impressions. As shown in the bottom panels of Fig. 1, during a simultaneous minutiae marking and eye-tracking task, examiners spend the vast majority of their time looking where they are marking. Thus, although our feature-marking task is only a proxy for the undisturbed behavior of a true fingerprint analysis, the eyes and marking behavior appear to be tightly coupled.

Following the marking task and no sooner than a week later, the same examiners conducted comparisons in which the regions they had previously selected were sequentially presented and individual decisions were made after each region was revealed. This allows us to track the growth of evidence throughout the comparison process and evaluate the role of relational information, feature diagnosticity, and image clarity.

The results were modeled using signal detection theory fit to individual subject data, which simultaneously characterizes the rates of information accumulation for identification and exclusion decisions.

Experiment 1

The goal of Experiment 1 was to determine the role of feature diagnosticity in the accumulation of evidence in a fingerprint comparison task. Sixteen fingerprint examiners selected eight regions from each of 72 high-quality impressions, using an interface as shown in the top panel of Fig. 2. This interface included only the latent impression, which mirrors casework because examiners typically mark up the latent impression before seeing candidate exemplar prints to avoid biases from the clear exemplar.

Participants moved a red square cursor around until they selected a region, which left a blue square. They were asked to select the most diagnostic region first, followed by seven additional diagnostic regions. The instructions emphasized choosing regions based on the utility of each region for purposes of comparison (which implied both identification and exclusion decisions).

At test, the pre-selected regions were sequentially presented to each examiner, and we asked the examiner to provide a tentative conclusion after each region was revealed. In addition, we revealed the regions in one of three orders:

- 1) Random (the diagnosticity rank of the regions was randomized)
- 2) Best to Worst Diagnosticity (the reveal of the regions followed the order in which they were selected)
- 3) Worst to Best Diagnosticity (the reveal of the regions was reversed relative to the order in which they were selected).

These orders allowed us to determine how diagnosticity might affect the manner in which evidence is accumulated.

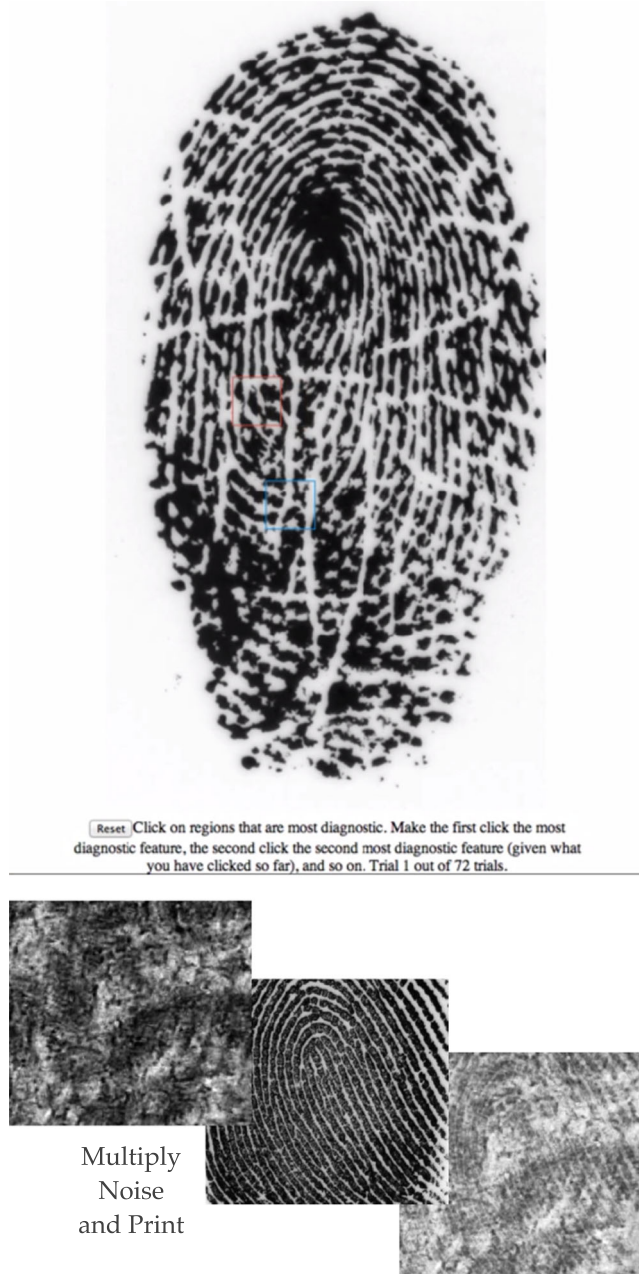


Fig. 2 **Top panel:** Illustration of the region selection screen. Participants used a red square to select regions for purposes of comparison in order of diagnosticity. The only constraint was that the regions could not physically overlap. One blue region has already been marked. **Bottom panel:** Noise combination mechanism. Left region – noise from black powder impression taken from regions with no ridge detail. Middle region – high-quality ink stamp impression. Right region – the result of a multiplicative process that combines the noise in ridge impressions to create a realistic simulated latent print

Method

Participants

Participants were 16 active fingerprint examiners (12 female) with at least 2 years of unsupervised casework experience.

These were recruited from state, federal, and large Metro labs in the USA. Participants were recruited from the International Association for Identification (IAI) annual conference, as well as regional associations of the IAI. Participants are therefore selected from a group who follows the IAI and are connected to the community of examiners through an association with the IAI. Due to the complexity of defining years of experience, we chose not to collect precise years of experience and only required our participants to be currently performing friction ridge examination casework and be 21 years or older. Other studies have not found strong links between years of experience and performance (Ulery et al., 2011, 2012).

Stimuli

The stimuli were collected in our laboratory using ink stamp pads on photo paper. This generally produces relatively high-quality impressions, and therefore for our comparison task we combined these images with noise patches sampled from black powder lifts that did not include ridge detail. The bottom panel of Fig. 2 illustrates an example of this process. Rather than use an additive noise process, we instead used a multiplicative process that treats the noise and ridge detail as neutral-density filters, which is appropriate for physical surfaces. This tends to create fairly realistic simulated latent prints (see the bottom panel of Fig. 2), and gives us control over the visibility of the regions. We will explicitly manipulated this in Experiment 2. No noise was used during the selection phase.

Procedure

The selection phase began with the following instructions:

In this experiment you will be using the mouse to indicate which regions of latent prints are likely to be most diagnostic. Think of this as the analysis phase of a comparison where you are identifying those regions that if you find them in the comparison print would provide the most diagnosticity or specificity for purposes of individualization. You will have 8 locations that you can click on, and each click will leave a square behind as a marker of where you clicked. You cannot overlap your clicks but you can put them next to each other. It is very important that your first click be on what you consider to be the most diagnostic feature or region on the print. Each subsequent click should be on regions that are progressively less diagnostic or provide progressively less specificity (but are still the most specific of the remaining areas).

The interface for this markup is shown in the top panel of Fig. 2. Participants completed 72 of these markups. Progress was self-paced, and typically extended over several days.

The comparison phase began at least 1 week after the end of the selection phase, and in some cases a month later. Examiners were shown displays similar to the top panel of Fig. 3 in which a single region was visible. After comparing this small region against the comparison print on the right side of the screen, the participants selected one of the eight decision options shown on the bottom of the display in the screenshots in Fig. 3, at which point a second region was revealed and participants were required to again select one of the eight options. The bottom panel of Fig. 3 illustrates a display from the same trial as in the top panel, but with all eight regions shown. Note that it is merely coincidental that there are eight total regions and eight possible choices for decisions.

We chose the response scale illustrated in Fig. 3 by combining difficulty with the traditional Exclusion/Identification language. In the psychological literature, the latent axis has been referred to as a confidence scale and the response labels adopt the language of confidence. However, we have found that fingerprint examiners greatly resist the conflation of their conclusion scale with the language of confidence. To ground their responses in a scale that they were familiar with, we adopted an expanded conclusion scale that mixed difficulty with the decision. However, the Inconclusive category was eliminated as a means to require fine judgments on the part of our participants. Because of these changes from a traditional scale (as well as our procedure of sequentially revealing small regions), our study should not be used to estimate error rates in the field of fingerprint examinations.

There were 72 total trials (36 mated and 36 non-mated) during the comparison phase of the experiment. No time limit was imposed on decisions, although the data were only saved at the end of each trial. Data collection typically did not take place in one sitting; it was often spaced out over several days or weeks. No feedback was given.

Results and discussion

The raw data for each subject are relatively straightforward: we get eight responses per trial, one for each region that is revealed. Each response is one of the eight possible choices ranging from Easy Exclusion to Easy Identification. Most participants generally initially chose Tending Exclusion or Tending Identification responses when only a single region was presented and then they selected responses toward the endpoints of the scale as additional regions were presented, although some reversals were observed.

The raw data consist of the counts at each of the eight possible responses for a given region count (e.g., five visible regions) accumulated across all trials for a given condition (Random, Best to Worst, or Worst to Best) for an individual subject.

One challenge with working with data along an Exclusion/Identification scale is that different examiners will have

different thresholds for how much evidence is required to make an “Identification” decision (this is known as a threshold for sufficiency for Identification in the latent print community). Where this threshold comes from and who gets to decide its value is a separate topic (see Mannering et al., 2021). However, for the present purposes we must acknowledge that these differences will exist and therefore transform our response count data into a value along an underlying evidence axis that represents the perceived strength of evidence for a given set of visible regions. Such an approach would separate the *strength* of the evidence from a *decision criterion* adopted by an examiner for the various decision options.

Modeling via signal detection theory

An obvious choice of model for this purpose is Signal Detection Theory (Macmillan & Creelman, 2005). Figure 4 illustrates how the distributions can be summarized using Gaussian curves, which are placed on a unitless axis as shown in Panel B of Fig. 4. We define the zero for this axis as the decision criterion between “Tending Exclude” and “Tending Ident” because this will allow us to simultaneously measure the evidence in favor of exclusion and the evidence in favor of identification. This fixed the endpoint of the latent dimension. The scale of the latent dimension is defined by the standard deviation of the distribution for data collected when four regions were present, which we fixed at 1.0 for all three conditions.² The fits to the other distributions (other than with four regions) have standard deviations that are free parameters, with the constraint that the mated and non-mated distributions have the same values for a particular number of regions. All three conditions (Random, Best to Worst, and Worst to Best) shared the same standard deviation values for a given region count. However, the means for the mated and non-mated distributions for these three conditions are allowed to vary as means to estimate the strength of evidence for a given condition and region count.

Panel B of Fig. 4 illustrates how we compute the locations of the mated and non-mated distributions, along with the distance between them, referred to as d' in the signal detection literature. Panel C of Fig. 4 illustrates how the evidence space is partitioned using the criteria that separate the different responses. Signal detection theory allows us to simultaneously estimate the decision criterion and the mated and non-mated distributions for all three region orderings and for each number of visible regions, subject to the following assumptions:

² We chose the data collected with four regions visible to represent a fixed standard deviation of 1.0 because we felt that these distributions are least likely to be affected by the endpoints of the scale or have data crowded into the “tending” responses in the middle of the scale.

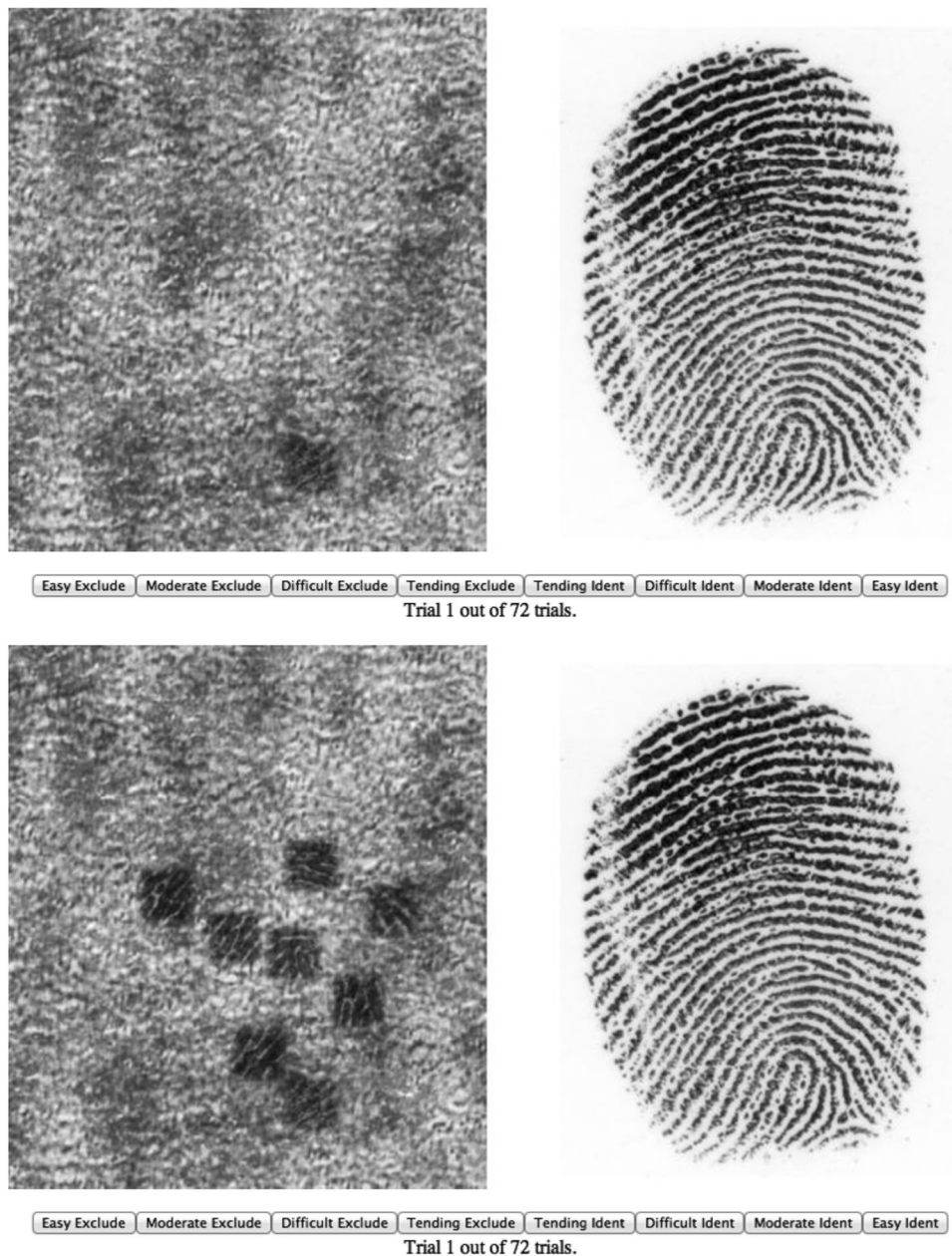


Fig. 3 **Top panel:** Example test trial with only a single region visible. Examiners make a response using one of the eight buttons after each region is presented. **Bottom panel:** Example test trial from the same sequence but with all eight regions visible. A trial requires eight

separate responses, one for each region, and the previous regions remain on the display as each new region is added. Because we have independent control over the noise and ridge impressions, the appearance of the new regions was quite natural

- a) We assume that subjects do not know which region ordering they are in (or even the existence of conditions; this was not revealed to them), and therefore we fit a common set of decision criterion (Panel C of Fig. 4) to the distributions for all three region orderings. We also applied the same set of decision criteria for all numbers of regions, with the assumption that each examiner would have one set of decision criteria that they would apply throughout the entire experiment, and for each number of visible regions.
- b) We assume that even if the collected data is not Gaussian (it is on an 8-point scale), the underlying evidence distributions are approximately Gaussian and the participant samples from these to compare against the decision thresholds and therefore generate a response. This Gaussian assumption has been tested in related memory paradigms with a 99-point scale and found to be approximately accurate (Mickes et al., 2007); therefore we are comfortable making this assumption (although see Rouder et al., 2007, and Wixted & Mickes, 2010) for a

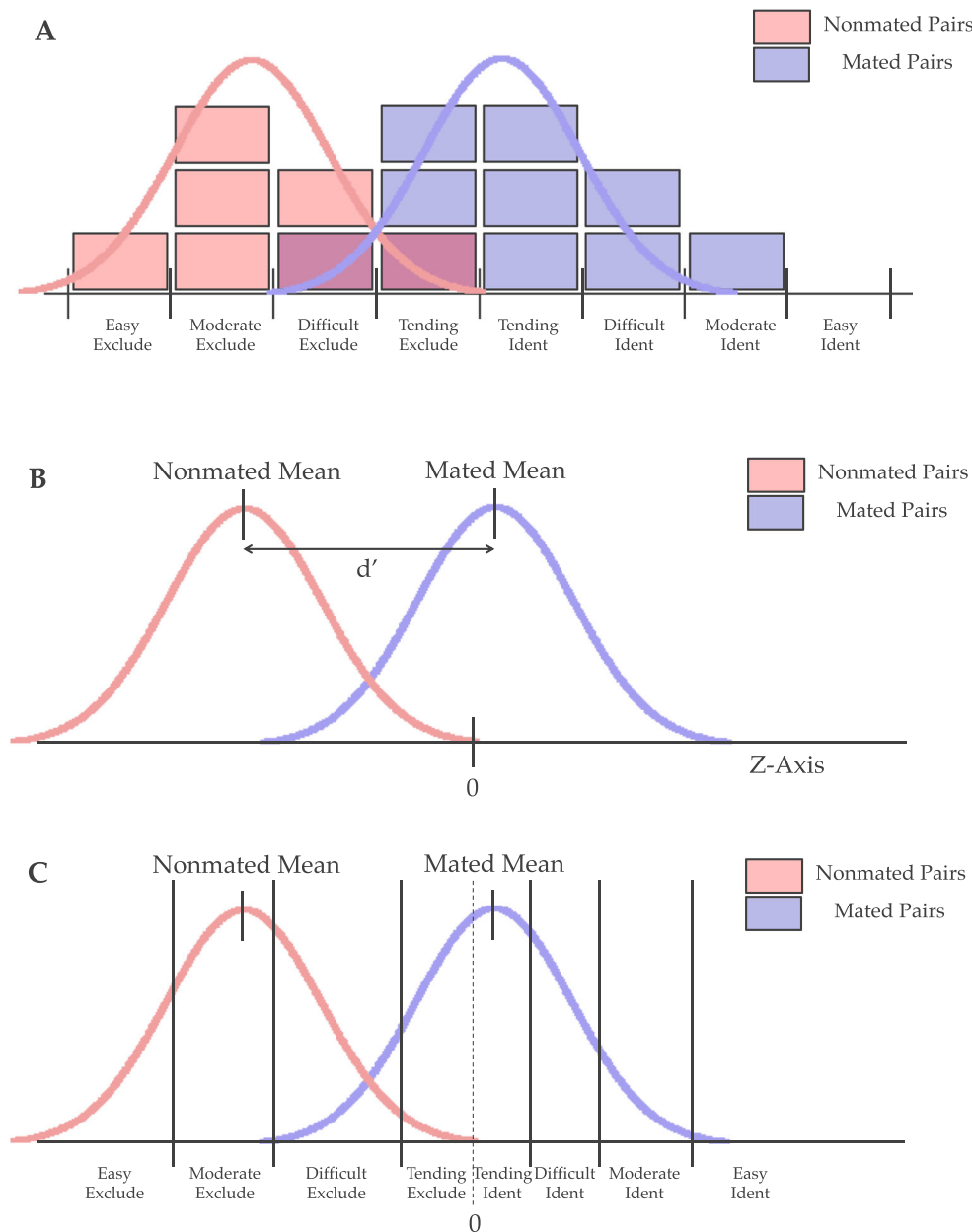


Fig. 4 Panel A: Summarizing the response distributions with Gaussian distributions. **Panel B:** Representing the locations of the two distributions using the non-mated and mated distribution means, which also measures d' . **Panel C:** All fitted parameters, including the six response criteria that partition the evidence axis into eight responses (the criterion that separates

Tending Exclusion from Tending Identification is fixed at zero to set the scale). Due to low number of Easy Exclude and Easy Identify responses, the actual modeling collapsed the “Easy Exclude” and “Easy Moderate” bins, as well as the “Easy Identify” and “Moderate Identify” bins

debate about whether the Gaussian assumption is testable). We primarily rely on the mean of the Gaussian, which is likely more stable than other attributes such as normality.

One challenge with fitting Signal Detection Theory to datasets with lots of possible responses is that some participants may fail to use a response category for some conditions. This was especially true for the “Easy Exclude” and “Easy Identify” responses, which produced relatively few responses for some

participants. Because we are mainly concerned with accurate estimates of the mated and non-mated distribution locations, we collapsed the “Easy Exclude” and “Easy Moderate” responses, as well as the “Easy Identify” and “Moderate Identify” responses. This produces a lower-dimensional parameter space and assists with the stability of the parameter estimation.

We use maximum likelihood estimation using custom Matlab code to estimate the parameters for each subject (see the Online Supplementary Material (OSM) via the Open Science Framework (OSF) at osf.io). For a complete dataset

for each participant, we estimate four decision criteria for the entire dataset (the boundary between Tending Exclude and Tending Ident was fixed at zero to fix the scale, and the extremes were collapsed, leaving just four criteria to be estimated). If we were to allow the criteria to shift when we added each additional patch, this would suggest that terms like “moderate ident” would take on a relative sense, such as “this is a moderate ident for a four-region image.” However, in casework the criteria for sufficiency are evaluated against some shared view of what is sufficient for identification. We believe that these external considerations provide the kind of constraints that would make the decision criteria stable across numbers of visible regions in our experiment. We estimated eight mated distribution locations, eight non-mated distribution locations, and seven standard deviation values for each region order condition (recall that the standard deviation for four patches was fixed at 1.0 to set the scale). These parameters are all estimated simultaneously using the `fMinSearch` function in Matlab. There are $4 + (8+8+7)*3 = 73$ free parameters, and the data contain $8*5*3 = 120$ degrees of freedom (eight separate decisions on each trial, five degrees of freedom for each number of visible regions, and three region orderings). Thus, the model is very far from saturation. If we had not collapsed the extreme responses (see previous paragraph), we would have had 75 free parameters and $8*7*3 = 168$ degrees of freedom.

Signal detection theory modeling results

The left column of Fig. 5 presents the results from the signal detection theory model, which in this implementation is a variant of ordinal or ordered probit regression. Model fits for individual participants are provided in a folder called `IndividualDPrimeGraphs` in each experiment folder on the OSF site, and while these exhibit considerable heterogeneity due to the fact that they are single-subject data, the general trends are consistent with the data presented in Fig. 5. The responses from each number of regions are modeled using a Gaussian distribution on the underlying latent axis with a mean and standard deviation that are estimated by choosing values that produce areas under the Gaussian distribution between the decision criterion that are similar to the observed frequencies from the participants. The four decision criteria that were allowed to vary are common to all conditions and numbers of visible regions. The adequacy of the model fits to individual subjects can be judged by inspection of the graphs in the folder `IndividualFrequenciesAndProportionsGraphs` on the OSF site. These graphs illustrate that the Gaussian distribution provides a reasonable summary of the response frequencies. Some of the mis-predictions of the model are likely due to multinomial variance. Note that we only rely on the mean of the Gaussian distribution, which is probably more robust than estimates of things like the normality of the data.

We summarize the discriminability of the examiners for each condition and number of regions by the difference between the mated and non-mated distributions (d' , see the top-left panel of Fig. 5) or by the location of the Gaussian distribution along the latent dimension (also referred to as the z-axis; see lower two panels of the left column of Fig. 5). Note that the two lower-left panels of Fig. 5 illustrate a conservative response bias, because the y-intercepts are negative. This likely results from the fact that examiners have unequal utilities: An erroneous identification can lead to dismissal, while an erroneous exclusion is not likely to even be caught.

As illustrated by the top-right panel of Fig. 5, sensitivity (d') grows approximately linearly as additional regions are added to the display. The growth of evidence is remarkably linear, with the exception of the Best to Worst region ordering, which shows a clear bend in both the mated and the non-mated distribution data and therefore the d' data as well (top panel). To determine whether there was a significant curvilinear component to the data, we fit both linear and quadratic regression models to the curves in Fig. 5, and used an F-test to determine whether the additional quadratic parameter produced a statistically significant improvement in the fit. All tests used an alpha of 0.05, which gives a critical F value of 6.61. For the d' fits in the top-left panel of Fig. 5, only the Best to Worst condition showed evidence of curvilinearity ($F(1,5) = 9.26$). For the mated trials in the middle-left panel of Fig. 5, again the Best to Worst condition showed evidence for curvilinearity ($F(1,5) = 77.88$). The Random condition in the non-mated trials showed evidence of curvilinearity ($F(1,5) = 10.45$) while the Best to Worst condition did not ($F(1,5) = 3.77$).

The data for the Worst to Best condition appears to show slightly worse performance than the other two conditions, a result that is conceptually replicated in Experiment 2. These two effects demonstrate small hysteresis effects: seeing a region with low diagnosticity early on appears to result in less willingness to use the endpoints of the conclusion scale when more regions are visible. Note that by the time all eight regions are visible in all three conditions, the images are identical in all three conditions, because presentation order is the only variable manipulated across the conditions.

The slope of the relation between the number of visible regions and the value of the mated or non-mated Gaussian distributions along the latent dimension can be viewed as a measure of the rate of information acquisition as more regions are added to the display. In addition to the apparent linearity of this relation, we can use the slopes of these lines as a rough guide to answer the following question: Is inculpatory information acquired faster or slower than exculpatory information? In principle, exclusions should be faster than identifications, because many examiners require only one unexplainable discrepancy to make an exclusion decision. However, identification decisions require exhaustive checking of all of

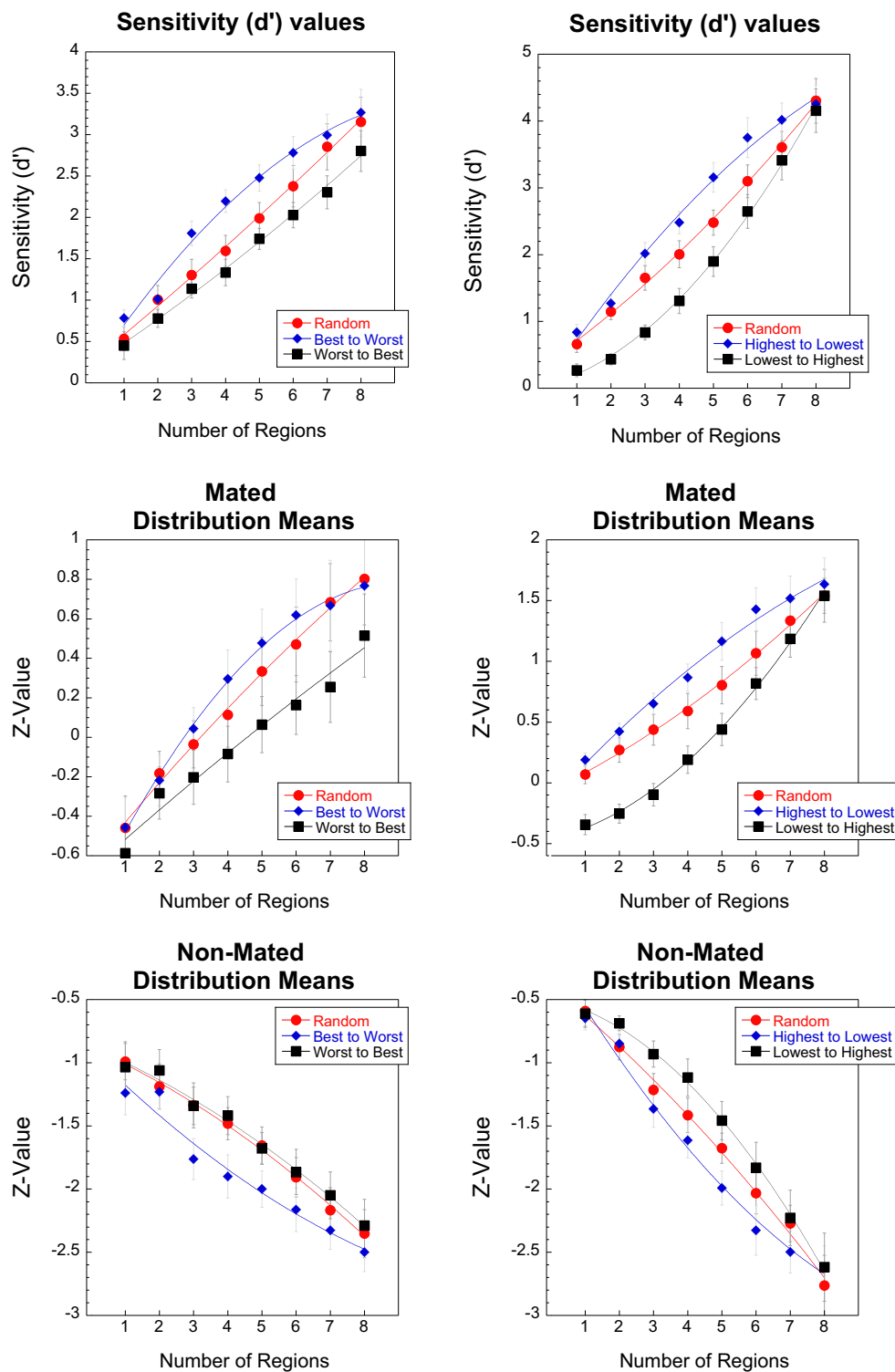


Fig. 5 **Left columns:** Data from Experiment 1 (manipulating region diagnosticity). **Right columns:** Data from Experiment 2 (manipulating image clarity). All error bars represent one standard error of the mean. **Top panels:** d' (sensitivity) data for all three region-order conditions. **Middle panels:** Location of the mated mean distribution along the evidence axis for all three region-order conditions. **Bottom panels:** Location

of the non-mated mean distribution along the evidence axis for all three region-order conditions. The growth of evidence for the random condition is mostly linear, while the Lowest to Highest region ordering shows a clear bend in both the mated and non-mated distribution data and therefore the d' data as well (top panels). Curves are quadratic regression lines. See text for details

the regions and therefore inculpatory information may grow at a slower rate.

Contrary to this expectation, the rate of growth of the two sets of curves in the lower-left two panels of Fig. 5 are quite similar (see also the left panel of Fig. 6). The slopes of the three mated conditions are 0.178, 0.178, and 0.139 for the Random, Best to Worst, and Worst to Best conditions. The slopes of the three non-mated conditions are -0.194, -0.186, and -0.185, and are negative because evidence accumulates in the exculpatory direction but are comparable to the mated condition by taking the absolute value of the slopes. Thus, exculpatory evidence may accumulate at a slightly faster rate, but this difference is small relative to the overall rate of information acquisition. The interpretation of this is somewhat complicated by the small but significant quadratic elements to the relations and should therefore be used only as a general guide to the differential rates of inculpatory and exculpatory information acquisition. This finding is conceptually replicated in Experiment 2.

Experiment 2

Feature diagnosticity is only one factor that might affect region utility. Region clarity has also been identified as a major factor in whether a region is marked or considered by examiners during a comparison (Ulery et al., 2014). What are the consequences of looking at low-clarity or high-clarity regions first during a comparison?

Experiment 2 is a conceptual replication of Experiment 1, and explores the role of feature clarity in the accumulation of evidence in a fingerprint comparison task. We asked a new set of 16 fingerprint examiners to select eight regions from each of 72 high-quality impressions, using an interface as shown in the top panel of Fig. 2. However, at test, we randomized the selection order (recall that Experiment 1 used Random, Best to Worst, and Worst to Best presentation orders), and instead experimentally manipulated the visibility of individual regions using a signal-to-noise manipulation. An example of this is shown in Fig. 7. We manipulated the signal-to-noise ratio (SNR) of the regions using the same multiplicative noise function used in Experiment 1. However, in Experiment 2 the SNR varied for each region by trading off the contrast of the noise and the fingerprint fragment in the multiplicative combination function.

We used this manipulation to reveal the regions in one of three orders:

- 1) Random (the visibility of the regions was randomized)
- 2) Highest Clarity to Lowest Clarity
- 3) Lowest Clarity to Highest Clarity.

These orders allow us to determine how feature visibility might affect the manner in which evidence is accumulated. In addition, this manipulation has another benefit, in that the visibility is much more visually apparent than feature diagnosticity might be, and we can also enforce a greater separation between the highest and lowest level. Note that in Experiment 1, feature diagnosticity might be fairly similar for all eight regions, in the sense that the top eight regions might all be of similar diagnosticity. However, we did observe differences in the presentation order in Fig. 5, so feature diagnosticity must have had some variation. Regardless, as shown in Fig. 7, the differences between the highest and lowest clarity regions are readily apparent in Experiment 2.

Method

Participants

Participants were 16 fingerprint examiners (eight female) with at least 2 years of unsupervised casework experience. These were recruited from the same population as in Experiment 1, be currently performing friction ridge examination casework, and be 21 years or older.

Stimuli

The stimuli were identical to Experiment 1, with the exception that the noise patches varied in their signal-to-noise level, which was done by increasing contrast of the visual noise while simultaneously decreasing the image contrast. We chose SNR levels that ranged from high to low clarity, and used the same selection procedure from Experiment 1 to generate personalized stimuli based on each participant's previous markup. The regions were combined with the noise using the same multiplicative combination mechanism, where image and noise contrast were constrained to sum to 1 as the SNR varied.

Procedure

The procedure was identical to Experiment 2, with the exception that the order of the regions followed one of the three orders described above (Random, Highest to Lowest Clarity, and Lowest to Highest Clarity).

Results and discussion

The results of the SDT modeling for Experiment 2 are shown in the right column of Fig. 5. As with Experiment 1, we see a fairly linear relation between d' and the number of visible regions (top-right panel of Fig. 5). However, there are quadratic components to all three conditions: Random ($F(1,5) =$

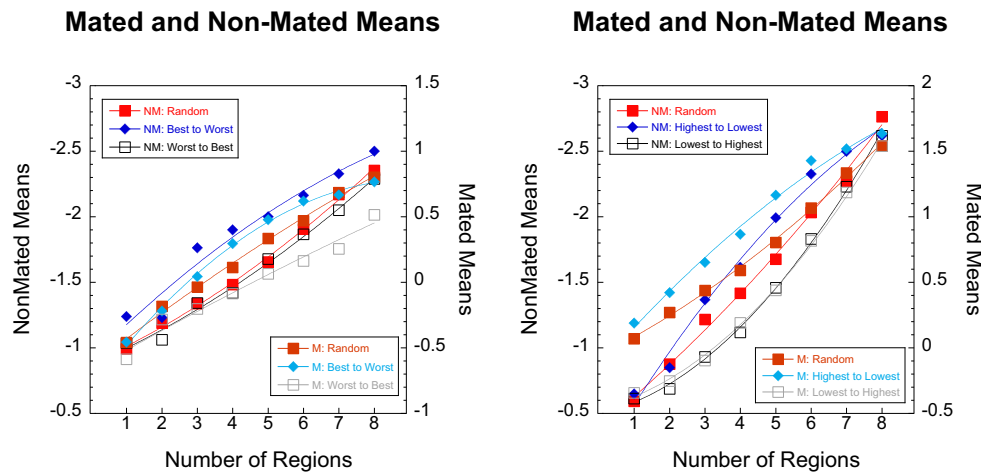


Fig. 6 Mated and non-mated data plotted on equivalently-scaled axes for Experiment 1 (left panel) and Experiment 2 (right panel). This demonstrates that the rate of evidence accumulation for non-mated pairs is fairly similar to that of mated pairs, with perhaps exculpatory evidence

11.29), Highest to Lowest ($F(1,5) = 7.25$), and Lowest to Highest ($F(1,5) = 102.19$). Of these, the Lowest to Highest is the most pronounced, implying that seeing low-clarity regions early on in the comparison can lead to low sensitivity but an accelerating acquisition rate to produce performance on par with the other two conditions by the time the eighth region is made visible.

For the mated trials, the relation between the number of regions and the location of the Gaussian distribution along the latent dimension is again mostly linear, with a small quadratic component for the Random condition ($F(1,5) = 13.75$) and a robust quadratic component for the Lowest to Highest ($F(1,5) = 67.45$) condition. The non-mated trials produced a significant quadratic term only for the Lowest to Highest condition ($F(1,5) = 81.52$).

Note that once all eight regions are visible, there is no physical difference between any of the three conditions, and it is noteworthy that the d' and z -axis values are virtually identical for the three conditions at the eight-region portion of the graph. Thus, any early effect of presenting high- or low-clarity regions early on the trial will dissipate by the point at which all eight regions are visible.

As with Experiment 1, we can get a rough sense of the rate at which inculpatory and exculpatory evidence accumulates by comparing the slopes of the mated and non-mated conditions, which fits linear regression functions and ignores any quadratic trends. These data are shown in the right panel of Fig. 6. The mated trials produce slopes of 0.211, 0.217, and 0.278 for the Random, Highest to Lowest, and Lowest to Highest conditions. The non-mated trials produce slopes of -0.296, -0.301 and -0.295 for the Random, Highest to Lowest, and Lowest to Highest conditions. Thus, it appears that exculpatory evidence accumulates at a slightly faster rate than inculpatory evidence, which is consistent with Experiment 1.

accumulating slightly faster than inculpatory evidence. Error bars are one SEM. Note that although all axes are different, they each span 2.5 standard deviation units and are therefore equivalent in span

In the modeling, we allowed the standard deviation of the Gaussian distribution on the latent dimension to be a free parameter, with the exception of the four-region presentation where it was fixed at 1.0 to set the scale. The same standard deviation was used for all three conditions for a given number of regions. As shown in Fig. 8, the standard deviations grew linearly as more regions were presented for both experiments, which simplifies the interpretation of the d' values and z -axis values in Fig. 5 because the underlying standard deviation is not changing in a non-linear way.

Inter-item dependencies

The modal finding of this work is a (mostly) linear relation between d' (sensitivity in signal detection theory (SDT) nomenclature) and the number of visible regions (see Fig. 5). This linear relation is consistent with an independent evidence accumulation model, as contrasted against a model in which a critical number of regions is required before they become self-reinforcing. However, there are important limitations to our analyses that must be acknowledged. Our SDT analysis groups together data from all trials for a given region count (e.g., five regions visible) and has no way to track the information growth on a *per trial basis*. If on some trials a subject was in a no-information state and then immediately transitioned into a full-information state at some particular region count (an “aha” moment, say with six regions), this would appear as an abrupt increase in d' at that region count. This clearly did not happen with regularity, because we saw (mostly) linear increases in d' with different region counts rather than an s-curve or step function. However, if this putative “aha” moment occurs early for some trials and later for other trials, this mixture could, in principle, average out to be a linear increase in d' with region count. We consider this unlikely – our curves

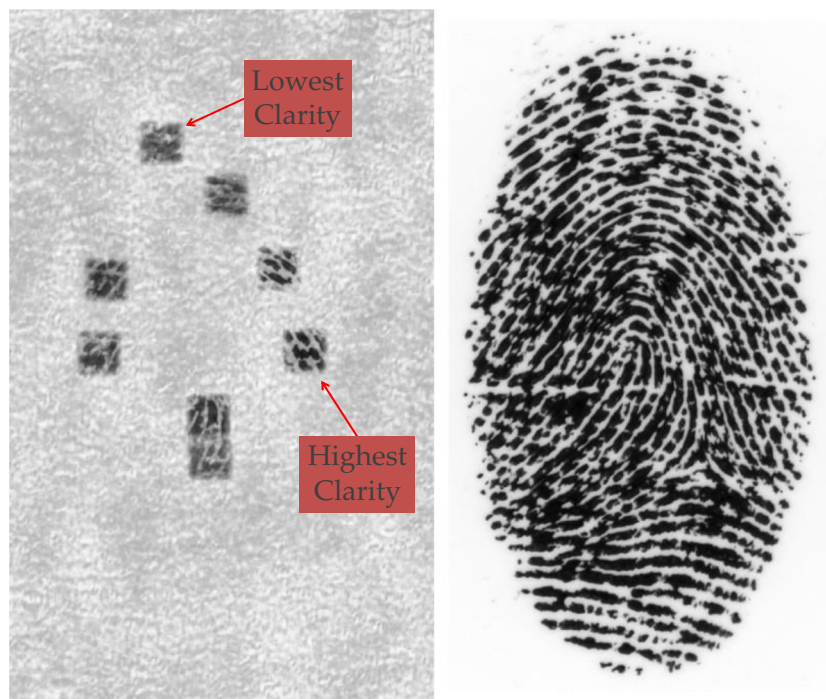


Fig. 7 Example manipulation of the task in Experiment 2. Rather than varying the diagnosticity of different regions, the diagnosticity order was randomized, and then regions were presented in one of three different clarity manipulations: Random, Highest to Lowest, or Lowest to Highest

show enough sensitivity to deviate from strict linearity with two separate manipulations – but we cannot rule this account out with our current approach. It would require a careful balancing of the mixture to produce a linear relation. However, without additional evidence we are reluctant to conclude that such “aha” events fail to occur, or that in some cases different regions can become self-reinforcing (Richler et al., 2015; Vogelsang et al., 2017). We would simply argue that, on the whole, such events are likely relatively rare and the typical behavior treats individual regions independently for purposes of evidence accumulation. However, there are other analyses that do consider the chain of responses within a trial, which we discuss next.

An alternative approach to addressing the question of whether evidence is accumulated gradually or in an all-or-none fashion relies on *transition probability matrices*. These are computed by conditioning on a given response with n regions and considering the likelihood of transitioning to a different response with $n+1$ regions. Summary transition probability matrix tables are shown in Table 1 for Experiment 1 for the random trials, with counts in each cell for each transition, and proportions in parentheses based on the row totals. We only analyzed the random trials because it is difficult to separate the physically diminishing information in the Best to Worst condition from a psychologically diminishing account where each region becomes psychologically less informative. The random condition does not systematically vary the informativeness, and so it is a cleaner condition to use

for the transition probability matrix. In principle, these tables should be considered separately as each region is added to the display, but for purposes of exposition we have collapsed across all of the responses within each trial and combined across subjects as well. The cells shaded gray in Table 1 correspond to cases where a region is added to the display but the examiner chose to use the same response as used previously. Green cells correspond to a one-step change toward an Identification conclusion, and blue toward an Exclusion conclusion. Yellow and orange cells represent the interesting cases, as they correspond to jumps of 2 or 3 steps respectively. Inspection of Table 1 illustrates relatively few larger jumps, with no cell representing more than 3% of the overall responses in that row, except in cases where there are almost no data in that row.

Table 2 illustrates the same analyses for Experiment 2. Again, we see very few counts in the yellow or orange shading, with the only exception being that 7% of the responses in the Tending ID response jump to Moderate ID for mated pairs. We also see 16% of the Difficult Exclusion response jump to Tending ID for mated pairs. Thus, we might describe this as only weak evidence for large transitions along the scale that would be consistent with a model where evidence accumulation was abrupt. Instead, the vast majority of the evidence is consistent with the conclusions from the SDT analysis: gradual accumulation of evidence with each patch contributing an independent amount of support for a given proposition.

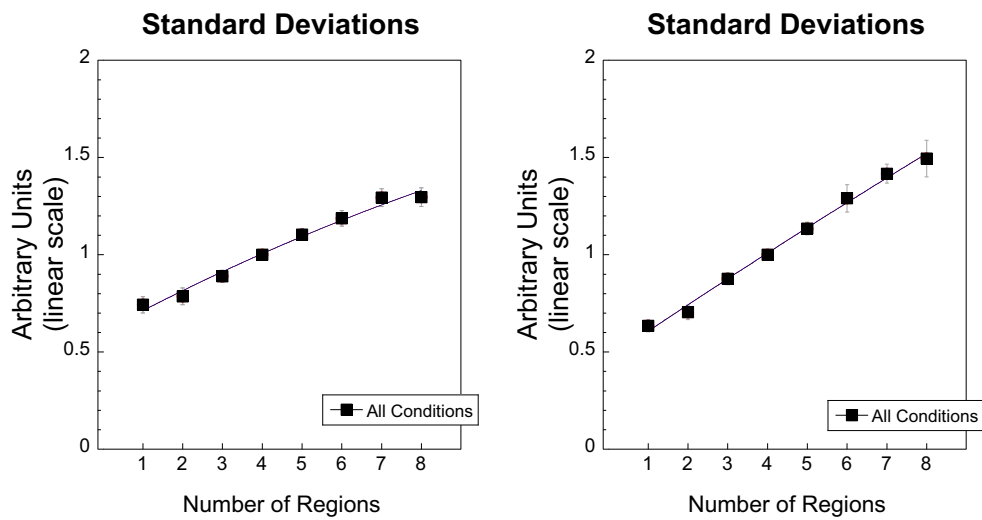


Fig. 8 Estimated standard deviation values for both Experiment 1 (left panel) and Experiment 2 (right panel) for different number of visible regions. Both functions are quite linear, which simplifies the interpretation of the graphs shown in Fig. 5. Error bars are one SEM

Table 1 Experiment 1 transition probability matrix between current response (row) and subsequent response (column) for non-mated (upper table) and mated trials (lower table) for the Random region ordering condition. Numbers in parentheses represent the proportion of responses for that row that fell in that cell

Non-Mated Impressions								
	EasyEx	ModEx	DiffEx	TendEx	TendID	DiffID	ModID	EasyID
EasyEx	106 (0.99)	0 (0.00)	0 (0.00)	1 (0.01)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
ModEx	27 (0.28)	67 (0.68)	2 (0.02)	2 (0.02)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
DiffEx	2 (0.02)	33 (0.28)	75 (0.63)	10 (0.08)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
TendEx	15 (0.01)	21 (0.02)	64 (0.06)	903 (0.87)	31 (0.03)	0 (0.00)	0 (0.00)	0 (0.00)
TendID	0 (0.00)	1 (0.01)	1 (0.01)	41 (0.28)	101 (0.70)	1 (0.01)	0 (0.00)	0 (0.00)
DiffID	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	1 (1.00)	0 (0.00)	0 (0.00)
ModID	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
EasyID	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
Mated Impressions								
	EasyEx	ModEx	DiffEx	TendEx	TendID	DiffID	ModID	EasyID
EasyEx	3 (0.75)	0 (0.00)	0 (0.00)	1 (0.25)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
ModEx	0 (0.00)	0 (0.00)	1 (0.33)	1 (0.33)	1 (0.33)	0 (0.00)	0 (0.00)	0 (0.00)
DiffEx	0 (0.00)	1 (0.06)	13 (0.76)	3 (0.18)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
TendEx	2 (0.00)	3 (0.01)	5 (0.01)	492 (0.84)	79 (0.13)	5 (0.01)	1 (0.00)	0 (0.00)
TendID	0 (0.00)	0 (0.00)	0 (0.00)	25 (0.06)	299 (0.74)	62 (0.15)	14 (0.03)	5 (0.01)
DiffID	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.01)	7 (0.05)	92 (0.65)	37 (0.26)	4 (0.03)
ModID	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.01)	2 (0.02)	49 (0.60)	30 (0.37)
EasyID	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	112 (1.00)

Color codes: Gray – stayed with same response when the next region is added. Green – advanced one response toward ID. Blue – advanced one response toward Ex. Yellow – advanced two responses either toward ID (upper triangle) or toward Ex (lower triangle). Orange – advanced three responses either toward ID (upper triangle) or toward Ex (lower triangle). Exclusions show some evidence for larger jumps (Yellow and Orange in lower triangle; upper table), but there is less evidence for large jumps for mated impressions (Yellow and Orange in upper triangle; lower table)

Table 2 Experiment 2 transition probability matrix between current response (row) and subsequent response (column) for non-mated (upper table) and mated trials (lower table) for the Random region ordering condition. Numbers in parentheses represent the proportion of responses for that row that fell in that cell

Non-Mated Impressions								
	EasyEx	ModEx	DiffEx	TendEx	TendID	DiffID	ModID	EasyID
EasyEx	97 (0.99)	0 (0.00)	0 (0.00)	1 (0.01)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
ModEx	27 (0.22)	96 (0.77)	1 (0.01)	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.01)	0 (0.00)
DiffEx	5 (0.03)	44 (0.22)	133 (0.67)	14 (0.07)	2 (0.01)	0 (0.00)	0 (0.00)	0 (0.00)
TendEx	6 (0.01)	28 (0.03)	94 (0.12)	648 (0.80)	37 (0.05)	1 (0.00)	0 (0.00)	0 (0.00)
TendID	0 (0.00)	2 (0.01)	4 (0.02)	59 (0.36)	96 (0.59)	1 (0.01)	0 (0.00)	0 (0.00)
DiffID	0 (0.00)	0 (0.00)	1 (0.50)	0 (0.00)	0 (0.00)	1 (0.50)	0 (0.00)	0 (0.00)
ModID	0 (0.00)	1 (1.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
EasyID	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)

Mated Impressions								
	EasyEx	ModEx	DiffEx	TendEx	TendID	DiffID	ModID	EasyID
EasyEx	0 (0.00)	0 (0.00)	1 (1.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
ModEx	1 (0.08)	6 (0.50)	2 (0.17)	1 (0.08)	1 (0.08)	0 (0.00)	1 (0.08)	0 (0.00)
DiffEx	0 (0.00)	7 (0.13)	35 (0.63)	4 (0.07)	9 (0.16)	1 (0.02)	0 (0.00)	0 (0.00)
TendEx	0 (0.00)	2 (0.01)	19 (0.05)	274 (0.73)	72 (0.19)	5 (0.01)	1 (0.00)	0 (0.00)
TendID	1 (0.00)	0 (0.00)	1 (0.00)	29 (0.07)	292 (0.68)	75 (0.17)	30 (0.07)	3 (0.01)
DiffID	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	4 (0.03)	93 (0.61)	52 (0.34)	4 (0.03)
ModID	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.01)	0 (0.00)	3 (0.02)	113 (0.70)	45 (0.28)
EasyID	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.01)	0 (0.00)	1 (0.01)	91 (0.98)

Color codes: Gray – stayed with same response when the next region is added. Green – advanced one response toward ID. Blue – advanced one response toward Ex. Yellow – advanced two responses either toward ID (upper triangle) or toward Ex (lower triangle). Orange – advanced three responses either toward ID (upper triangle) or toward Ex (lower triangle). Mated and non-mated trials show weak evidence for larger jumps (Yellow and Orange)

General discussion

It is perhaps somewhat surprising that support for the same source and different sources propositions accumulates at approximately the same rate. This is contrary to expectations in the examiner community, who often rely on “one unexplainable difference” when making exclusion decisions. We also found relatively little support for any form of sequential biasing or hysteresis in the sequential decisions (Kiyonaga et al., 2017) because even though information accumulation was lessened by poor quality patches presented early in the trial, the presentation of higher-quality patches later in the trial made up for this slow start. This is perhaps good news for the examiner community, who may not have to rely on always finding the most diagnostic region first or risk biasing the entire decision process for a given comparison.

The independence we observe between patches may be a function of their relative spatial isolation. We know that these are relevant patches, because each examiner selected their own regions. However, while fingerprints often contain regions of high and low quality similar to our displays, they are often connected by at least discernable ridge flow. Had we included such background features, we may have observed one region “bootstrapping” nearby regions, thus producing d’ curves that accelerated rather than appearing linear with the number of presented regions. However, designing such a background while not providing additional information might prove challenging.

The results help constrain models of spatial information acquisition, because complex models that involve integration across regions may not be necessary to account for the major finding that items seem to be processed and interpreted

relatively independently. This result may be region-size dependent, and we had to make some decisions about how large our regions were based on the relative size of minutiae in our impressions. Although independence may break with extremely small or extremely large patches, independence seems to hold for our mid-sized regions.

Open practices statement The data and materials for all experiments are available via the Open Science Framework at osf.io. The project was not preregistered because data collection began before this became a standard practice.

References

- Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, 93(2), 154–179. <https://doi.org/10.1037/0033-295x.93.2.154>
- Busey, T., & Vanderkolk, J. R. (2005). Behavioral and electrophysiological evidence for configural processing in fingerprint experts. *Vision Research*, 45(4), 431–448.
- Busey, T., Yu, C., Wyatte, D., & Vanderkolk, J. (2013). Temporal sequences quantify the contributions of individual fixations in complex perceptual matching tasks. *Cognitive Science*, 37(4), 731–756. <https://doi.org/10.1111/cogs.12029>
- Busey, T., Swofford, H. J., Vanderkolk, J., & Emerick, B. (2015). The impact of fatigue on latent print examinations as revealed by behavioral and eye gaze testing. *Forensic Science International*, 251, 202–208. <https://doi.org/10.1016/j.forsciint.2015.03.028>
- Busey, T., Nikolov, D., Yu, C., Emerick, B., & Vanderkolk, J. (2017). Characterizing human expertise using computational metrics of feature diagnosticity in a pattern matching task. *Cognitive Science*, 41(7), 1716–1759. <https://doi.org/10.1111/cogs.12452>
- Busey, T., Heise, N., Hicklin, R. A., Ulery, B. T., & Buscaglia, J. (2021). Characterizing missed identifications and errors in latent fingerprint comparisons using eye-tracking data. *PLoS One*, 16(5), e0251674.
- Fific, M., & Townsend, J. T. (2010). Information-processing alternatives to holistic perception: Identifying the mechanisms of secondary-level holism within a categorization paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(5), 1290.
- Hicklin, R. A., Ulery, B. T., Busey, T. A., Roberts, M. A., & Buscaglia, J. (2019). Gaze behavior and cognitive states during fingerprint target group localization. *Cognitive Research-Principles and Implications*, 4, 12. <https://doi.org/10.1186/s41235-019-0160-9>
- Holland, M. K., & Lockhead, G. R. (1968). Sequential effects in absolute judgments of loudness. *Perception & Psychophysics*, 3(6), 409–414. <https://doi.org/10.3758/Bf03205747>
- Kim, B., Reif, E., Wattenberg, M., Bengio, S., & Mozer, M. C. (2021). Neural networks trained on natural scenes exhibit gestalt closure. *Computational Brain & Behavior*, 4(3), 251–263. <https://doi.org/10.1007/s42113-021-00100-7>
- Kiyonaga, A., Scimeca, J. M., Bliss, D. P., & Whitney, D. (2017). Serial dependence across perception, attention, and memory. *Trends in Cognitive Sciences*, 21(7), 493–497.
- Kondo, A., Takahashi, K., & Watanabe, K. (2012). Sequential effects in face-attractiveness judgment. *Perception*, 41(1), 43–49. <https://doi.org/10.1068/p7116>
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: a user's guide* (2nd ed.). Lawrence Erlbaum Associates.
- Mannering, W. M., Vogelsang, M. D., Busey, T. A., & Mannering, F. L. (2021). Are forensic scientists too risk averse? *Journal of Forensic Sciences*, 66(4), 1377–1400. <https://doi.org/10.1111/1556-4029.14700>
- Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin & Review*, 14(5), 858–865. <https://doi.org/10.3758/Bf03194112>
- Mochon, D., & Frederick, S. (2013). Anchoring in sequential judgments preface. *Organizational Behavior and Human Decision Processes*, 122(1), 69–79. <https://doi.org/10.1016/j.obhdp.2013.04.002>
- Piepers, D. W., & Robbins, R. A. (2012). A review and clarification of the terms "holistic," "configural," and "relational" in the face perception literature. *Frontiers in Psychology*, 3, 559. <https://doi.org/10.3389/fpsyg.2012.00559>
- Richler, J. J., Palmeri, T. J., & Gauthier, I. (2012). Meanings, mechanisms, and measures of holistic processing. *Frontiers in Psychology*, 3, 553. <https://doi.org/10.3389/fpsyg.2012.00553>
- Richler, J. J., Palmeri, T. J., & Gauthier, I. (2015). Holistic processing does not require configural variability. *Psychonomic Bulletin & Review*, 22(4), 974–979. <https://doi.org/10.3758/s13423-014-0756-5>
- Rouder, J., Pratte, M., & Morey, R. (2007). Latent mnemonic strengths are latent: A comment on Mickes. *Wixted and Wais*, 17(3), 427–435.
- Searston, R. A., & Tangen, J. M. (2017). Expertise with unfamiliar objects is flexible to changes in task but not changes in class. *PLoS One*, 12(6), e0178403. <https://doi.org/10.1371/journal.pone.0178403>
- SWGFAST. (2013). Document #10 Standards for Examining Friction Ridge Impressions and Resulting Conclusions (Latent/Tenprint). Retrieved from https://www.nist.gov/system/files/documents/2016/10/26/swgfast_examinations-conclusions_2.0_130427.pdf
- Swofford, H. J., & Cino, J. G. (2017). Lay understanding of "identification": How jurors interpret forensic identification testimony. *Journal of Forensic Identification*, 68(1), 29–41.
- Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology Section a-Human Experimental Psychology*, 46(2), 225–245. <https://doi.org/10.1080/14640749308401045>
- Taylor, M., Chapman, W., Hicklin, A., Kiebuszinski, G., Mayer-Splain, J., Wallner, R., & Komarinski, P. (2013). *Extended feature set profile specification*. National Institute of Standards and Technology, Gaithersburg, MD, NIST SP 1134.
- Thompson, P. (1980). Thatcher, Margaret - a new illusion. *Perception*, 9(4), 483–484. <https://doi.org/10.1068/p090483>
- Thompson, M. B., Tangen, J. M., & Searston, R. A. (2014). Understanding expertise and non-analytic cognition in fingerprint discriminations made by humans. *Frontiers in Psychology*, 5, 737. <https://doi.org/10.3389/fpsyg.2014.00737>
- Tierney, L. (2013). Analysis, comparison, evaluation, and verification (ACE-V). In Houck, M., & Tenney, S. (Eds.), *Forensic Fingerprints* (3rd ed., pp. 73–74). Elsevier, 2016.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty - heuristics and biases. *Science*, 185(4157), 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>
- Ulery, B. T., Hicklin, R. A., Buscaglia, J., & Roberts, M. A. (2011). Accuracy and reliability of forensic latent fingerprint decisions. *Proceedings of the National Academy of Sciences of the United States of America*, 108(19), 7733–7738. <https://doi.org/10.1073/Pnas.1018707108>
- Ulery, B. T., Hicklin, R. A., Buscaglia, J., & Roberts, M. A. (2012). Repeatability and reproducibility of decisions by latent fingerprint examiners. *PLoS One*, 7(3), e32800. <https://doi.org/10.1371/journal.pone.0032800>
- Ulery, B. T., Hicklin, R. A., Roberts, M. A., & Buscaglia, J. (2014). Measuring what latent fingerprint examiners consider sufficient information for individualization determinations. *PLoS One*, 9(11), e110179. <https://doi.org/10.1371/journal.pone.0110179>

- Vogelsang, M. D., Palmeri, T. J., & Busey, T. A. (2017). Holistic processing of fingerprints by expert forensic examiners. *Cognitive Research: Principles and Implications*, 2(1), 15. <https://doi.org/10.1186/s41235-017-0051-x>
- Wixted, J. T., & Mickes, L. (2010). Useful scientific theories are useful: A reply to Rouder, Pratte, and Morey (2010). *Psychonomic Bulletin & Review*, 17(3), 436–442.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.