

Article

Not peer-reviewed version

Converging Information-Theoretic Measures of Metacognitive Efficiency

[Daniel Fitousi](#) *

Posted Date: 5 February 2025

doi: 10.20944/preprints202502.0274.v1

Keywords: Metacognition; Confidence; Metacognitive Efficiency; Mutual Information; KL Distance; 17 Jeffrey' Distance



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Article

Converging Information-Theoretic Measures of Metacognitive Efficiency

Daniel Fitousi 

Department of Psychology, Ariel University, Israel; daniel.fi@ariel.ac.il

Abstract: The ability of participants to monitor the correctness of their own decisions by rating their confidence is a form of metacognition. This introspective act is crucial for many aspects of cognition including perception, memory, learning, emotion regulation, and social interaction. Researchers assess the quality of confidence ratings according to *bias*, *sensitivity*, and *efficiency*. To do so, they deploy such quantities as $meta - d' - d'$, or $M - ratio$ [1,2]. These measures compute the expected accuracy level of performance in the primary task (Type 1) from the secondary confidence rating task (Type 2). However, these measures have several limitations. For example, they rely on unwarranted parametric assumptions, and they fall short of accommodating the granularity of confidence ratings. Two recent papers by Dayan [3] and Fitousi [4] have proposed information-theoretic measures of metacognitive efficiency that can ameliorate these problems. Dayan suggested $meta - \mathcal{I}$, and Fitousi proposed: $meta - \mathcal{U}$, $meta - \mathcal{KL}$, and $meta - \mathcal{J}$. These authors demonstrated the convergence of their measures on the notion of metacognitive efficiency using simulations, but did not apply their measures to real empirical data. The present study set to test the convergence of these measures in a concrete behavioral task using two independent data sets. The present results supported the viability of these novel indexes of metacognitive efficiency, and provide substantial empirical evidence for their convergence.

Keywords: metacognition; confidence; metacognitive efficiency; mutual information; KL distance; Jeffrey' distance

1. Introduction

Assessing the validity of our own decisions is a self-reflective act of immense survival value. For example, while learning for a test, we have to constantly monitor the level of our knowledge. If we feel that a satisfactory level of knowledge has been accomplished we can halt our preparation, otherwise we have to continue our learning [4,5]. This self-reflective ability is a form of *metacognition* [6–12]. Metacognition plays a central role in various domains, including: learning [13], memory [14,15], self-awareness [16], action [17,18], psychiatric disorders [19], and social interaction [20]. Several neural substrates are involved in metacognitive assessments and monitoring of behavior [22]. These include the right anterior dorsolateral prefrontal cortex and the bilateral parahippocampal cortex [10]. A prominent method of assessing metacognition behaviorally is based on collecting confidence judgments from participants [12]. Measuring the quality of these confidence ratings is consequential for understanding the underlying mechanisms of metacognition. The goals of the present study are threefold: (a) to introduce several recently developed information-theoretic measures of metacognitive efficiency that are based on confidence judgments [3,4], (b) to demonstrate how these measures can be applied to a concrete behavioral task of face-matching [21], and (c) to show that these information-theoretic measures converge on a single theoretical construct in this empirical context [22].

A primary objective of research on metacognition is that of measuring the quality of confidence judgments. Three quantities stand out as crucial aspects of measurement and modeling: *bias*, *sensitivity*, and *efficiency*. To motivate our understanding of these concepts, consider a simple task by which an observer is asked to detect whether a signal is present or absent by emitting one of two possible responses, and then immediately she is asked to rate her confidence in the correctness of her decision.

It is often customary to call the first task *type 1* task, and the second *type 2* task [23,24]. It is also theoretically convenient to separate the actor, who performs in the type 1 task, from the rater, who engages in the type 2 task, although the two entities might be the same person [25]. Sensitivity plays a role in both the actor's and the rater's decisions. The actor's sensitivity is reflected in her ability to separate between noise and noise plus signal. In the language of Signal Detection Theory [SDT, 26,27] this quantity is captured by the parameter d' . The rater's sensitivity reflects his ability to distinguish between the actor's correct and incorrect decisions. A metacognitively sensitive rater tends to provide high confidence ratings for the actor's correct decisions, and low confidence ratings for the actor's incorrect decisions. However, the sensitivity of the rater cannot serve as a viable measure of the rater's quality of decision. Trivially, a rater can be highly sensitive if the actor is giving correct decisions most of the time. In other words, the rater's job becomes much easier as the actor becomes more sensitive. Thus, a genuine measure of metacognitive efficiency should correct the rater's sensitivity for the actor's performance. Moreover, metacognitive efficiency can be affected by bias. An overconfident rater, one who gives high confidence ratings even when the actor's decisions are incorrect, is effectively both insensitive and inefficient.

Model-free [28] and model-based [1,2,24,25,29] approaches have been developed to quantify metacognitive bias, sensitivity and efficiency. One of the most influential model-based frameworks in the field is called the *meta* - d' model [1,2]. This model (see Figure 1) relies on principles of SDT [26,27], and as such incorporates parametric assumptions [23,24,30]. The model's objective is to predict the rater's hypothetical sensitivity in the type 1 task based on the rater's performance in the type 2 task. The value of this theoretical quantity is dubbed *meta* - d' . Because this quantity and the actor's type 1 d' are expressed in the same measurement units [2], they can be directly compared. Metacognitive efficiency can then be assessed in two possible ways: either as an M-ratio (*meta* - d' / d') or as a difference (*meta* - d' - d'). The rater's efficiency is optimal when *M* - ratio = 1, or when *meta* - d' - d' = 0. The rater is hyposensitive when *M* - ratio < 1 or when *meta* - d' - d' < 0. There are situations in which the rater is hypersensitive, namely when *M* - ratio > 1, or when *meta* - d' - d' > 0 [1]. This can occur when the rater has access to additional information accumulated after the actor's decision. Such a scenario can be modeled, for example, by the second-order model [25]. According to this model, the actor and the rater are sampling evidence from two separate distributions. Hypersensitivity in this model can be modeled by making the noise (i.e., variance) in the rater's distribution much lower than the noise (i.e., variance) in the actor's distribution.

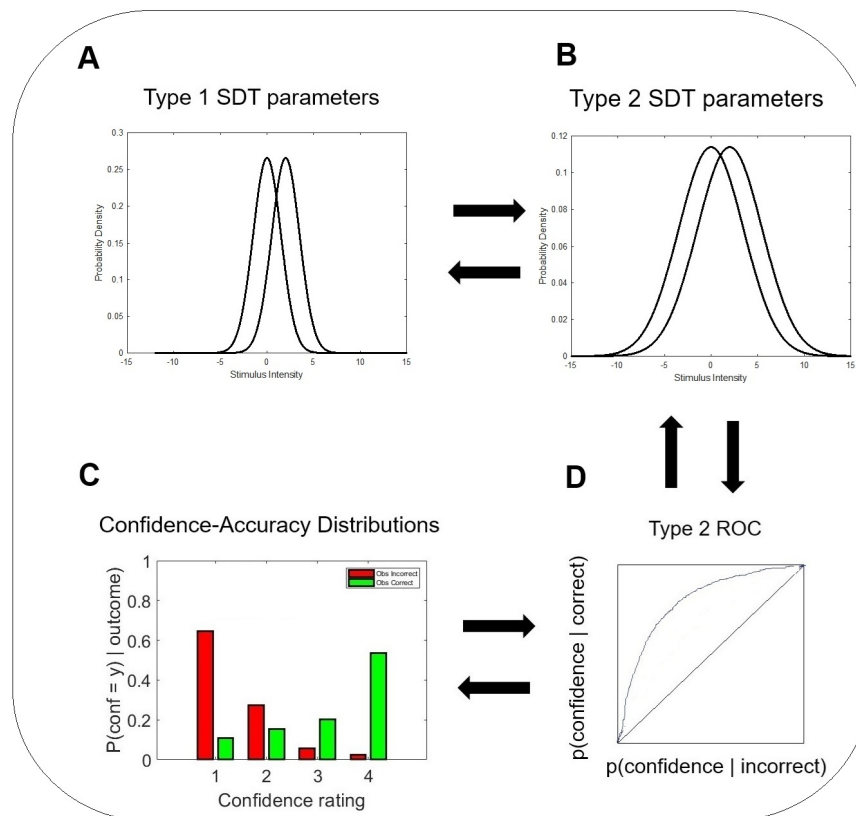


Figure 1. The meta - d' model [1]. (A) A signal detection theory (SDT) representation of the actor's decisions in the type 1 task. (B) An SDT representation of the rater's decisions in the type 2 task. (C) Empirical confidence-rating distributions conditional on type 1 correct and incorrect responses. A rater with high metacognitive sensitivity provides high confidence ratings when the actor is correct, and low confidence ratings when the actor is incorrect. A rater with low metacognitive sensitivity does not distinguish between the actor's correct and incorrect responses. (D) A theoretical ROC curve of the type 2 task is generated by assuming a hypothetical equal-variance SDT model of confidence. The meta- d' model [1,29,31] expresses the type 2 hit rates $p(Hit)$ and false alarms $p(FA)$ in terms of the divergence units of the type 1 d' , and then looks for the type 1 d' that best fits the empirical confidence data. The resulting parameter is called *meta - d'* .

Several measures emanating from the *meta - d'* model [1,2], such as *meta - d' - d'* and *M - ratio*, have been used extensively by researchers to evaluate the quality of confidence judgments [32]. However, these measures are not without their limitations [3]. First, it is not immediately clear why empirical confidence ratings should necessarily accommodate to a type 1 decision process [3]. Second, strong evidence suggests that these measures carry-over a residual dependency on type 1 performance [33], and on metacognitive bias [34]. Beyond these two reasons, these measures: (a) require distributional assumptions, (b) are not bias-free, (c) are applicable only to ratio-scale variables, (d) do not scale with the number of stimuli, responses, or confidence rating alternatives, (e) do not have definite upper and lower bounds, and (f) are applicable only to experimental designs with two stimuli and two responses. Dayan [3] and Fitousi [4] have recently proposed information-theoretic indexes of metacognitive efficiency that can ameliorate most of these problems, and serve as additional means of assessment.

Dayan [3] has succinctly summarized the utility of an information-theoretic approach to judgments of confidence: "the problem of confidence is inherently one of information – that the actor has about the true state of the stimulus; and that the rater has about the same quantity and about what the actor used. It therefore seems appropriate to use the methods of information theory to judge the relationship between the stimulus, the actor and the rater" (p.11). Indeed, the information theoretic measures have many desirable characteristics: (a) they are relative measures, (b) accommodate to type 1 performance, (c) bias- free (e.g., do not depend on criterion shifts), (d) do not require any

parametric assumptions, (e) scale with the number of confidence rating points, (f) do not require the assumption of ratio or order of confidence ratings, and (g) relatively easy to compute. Both Dayan [3] and Fitousi [4] performed simulation with canonical Bayesian models to support the validity of the information-theoretic measures they developed. Both authors demonstrated convergence of their information-theoretic measures based on those SDT and related Bayesian models. However, neither of them computed the measures in real empirical data sets, nor did they examine the convergence validity of those measures in such cases. The present study sought to accomplish these objectives.

2. An information-Theoretic Approach to Metacognitive Efficiency

Recent years have seen a renewed interest [35] in information-theoretic approaches to perception and cognition [36–44]. Many researchers have begun to realize the potential strength of information theory [45,46] in providing powerful tools for addressing the uncertainty that characterizes brain and behavior [47,48]. In two recent studies, Dayan [3] and Fitousi [4] have developed complementary information-theoretic approaches to metacognition that build on the idea of confidence judgments as a communication system. An illustration of such a system is presented in Figure 2. The communication model consists of three entities: Stimuli which are set by the Experimenter, Actor who responds to the stimuli, and Rater who provides confidence judgments of the Actor's response to the Stimuli.

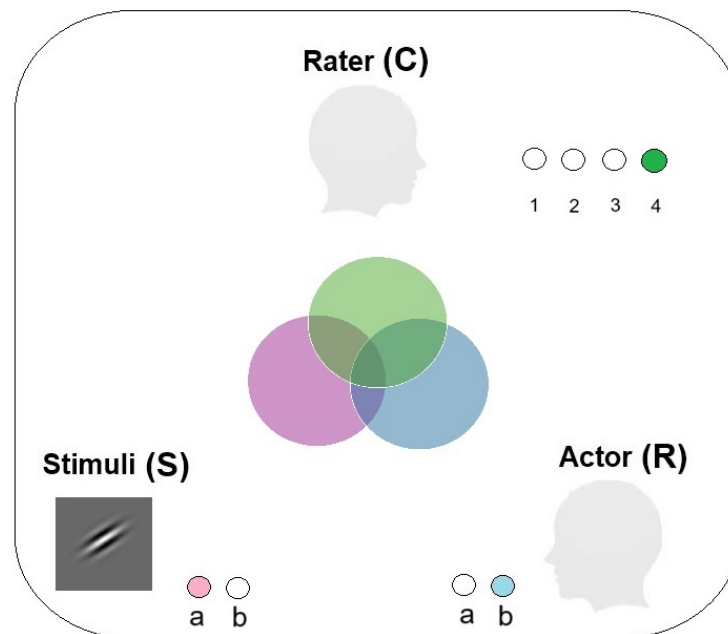


Figure 2. A Venn diagram illustrating a communication system between stimuli, actor and rater. Uncertainty in Stimuli is depicted by a pink circle, uncertainty in Actor's response is illustrated by a blue circle, and uncertainty in Rater's confidence ratings is represented by a green circle. In this example, there are two stimuli. The actor responds by emitting one of two possible responses, and the rater expresses their confidence in the actor's response by sending one of four confidence-rating levels.

2.1. Dayan's $Meta - \mathcal{I}$

Dayan [3] outlined an information-theoretic measure of metacognitive efficiency he called $meta - \mathcal{I}$. Consider an actor who performs in a two-choice signal detection task with consequences of a =1, or a = -1. The actor's response is r , and she is correct when $r = a$ and incorrect when $r \neq a$. The confidence-accuracy discrete distributions can be used to assess the metacognitive efficiency of the rater as a difference between two entropies.

$$meta - \mathcal{I} = H_2(r|a = -1) - H_2(r|c, a = -1) \quad (1)$$

where

$$H_2(r|a = -1) = h_2[P(r|a = -1)] \quad (2)$$

and

$$H_2(r|c, a = -1) = \sum_c P(c|a = -1) h_2[P(r|c, a = -1)] \quad (3)$$

where

$$h_2[P(x)] = \sum_{x=0;1} -P(x) \log_2 P(x) \quad (4)$$

is the entropy of a Bernoulli random variable. The first entropy $H_2(r|a = -1)$ is the overall uncertainty about the actor's accuracy. In the SDT model it is determined by the value of the parameter d' . When the actor is perfectly accurate and $d' \rightarrow \infty$ or when the actor is perfectly wrong, this entropy amounts to 0 bits; when the actor is completely uncorrelated with d' , which occurs when $d' \rightarrow 0$, this entropy amounts to 1 bits. The second entropy $H_2(r|c, a = -1)$ is the weighted average uncertainty about the residual accuracy after observing the confidence rating c where the weights arrive from the probability of observing that rating c .

The *meta* – $\mathcal{I}$ measure has many desirable characteristics: (a) it is a relative measure, (b) accommodates to type 1 performance, (c) does not require any parametric assumptions, (d) is bias-free, (e) scales with the number of confidence rating points, (f) does not require the assumption of ratio or order of confidence ratings, and (g) has definite bounds from above and from below.

The *meta* – $\mathcal{I}$ is an ingenious measure on its right. But it can be modified to become a more general index of metacognitive efficiency. Note that the measure is contingent on accuracy (i.e., it depends on a specific partition, $r = a$, $r \neq a$), while there is strong evidence that participants do not necessarily rate confidence based on the accuracy of response [49]. For example, Koriati [52,53] has demonstrated the operation of a *consensuality principle* in judgments of confidence, whereby participants give the highest confidence rating to the most popular option, not necessarily to the one that is objectively correct. Moreover, *meta* – $\mathcal{I}$ was originally developed to assess metacognitive efficiency in the two-choice SDT task (i.e., it assumes two mutual choices, $a = 1$ or $a = -1$). However, *meta* – $\mathcal{I}$ does not apply to tasks of multi-alternative choice [49]. For example, assume an experimental design with four stimuli that are categorized into four responses, and are then rated on a 1-4 confidence scale. In that case, *meta* – $\mathcal{I}$ is not readily applicable. Thus, a more general version of *meta* – $\mathcal{I}$, I call *meta* – $\mathcal{I}'$, is needed. This version is not contingent on accuracy, and makes no assumptions about the structure of the experimental design (i.e., number of stimuli or/and responses). The new *meta* – $\mathcal{I}'$ proposed here meets these requirements. It is computed as the mutual information between response and the joint confidence-stimuli variable:

$$\begin{aligned} \text{meta} - \mathcal{I}' &= I(r; c, s) = H(r|s) - H(r|s, c) \\ &= H(r) + H(c, s) - H(s, r, c) \end{aligned} \quad (5)$$

This quantity amounts to 0 bits when the actor shares no information with the stimuli and rater; it is maximal when the actor shares all the information with the stimuli and the rater. An upper-bound on *meta* – $\mathcal{I}'$ is determined by the variable with the lowest number of alternatives. For example, if the stimuli variable has the lowest number of alternatives, say four, then the upper bound is 2 bits. In the case of a two-choice SDT task, the upper bound is 1 bit. Decomposition of *meta* – $\mathcal{I}'$ into its entropy terms can provide more insights. According to one possible decomposition, it reflects the degree to which uncertainty in the actor's response to the stimuli $H(r|s)$ is reduced after observing the rater's confidence judgments $H(r|s, c)$. According to another decomposition, *meta* – $\mathcal{I}'$ reflects the reduction in the combined uncertainty in the actor's response and the confidence-stimuli variables $H(r) + H(s, c)$ after removing the uncertainty in the joint stimuli-response-confidence variable $H(s, r, c)$. In its new

form, $meta - \mathcal{I}'$ can be applied to complex experimental designs, with several stimuli and responses, and without the need to define the correctness of the response.

2.2. Fitousi's $meta - \mathcal{U}$, $meta - \mathcal{KL}$, and $meta - \mathcal{J}$

Fitousi [4] has recently developed three complementary information-theoretic measures of metacognitive efficiency he called: $meta - \mathcal{U}$, $meta - \mathcal{KL}$, and $meta - \mathcal{J}$. The first measure bears close affinity with Dayan's $meta - \mathcal{I}$, and its generalized form $meta - \mathcal{I}'$. The measure quantifies metacognitive efficiency as the mutual information exchanged between the rater's confidence judgments and the stimuli-response variable:

$$\begin{aligned} meta - \mathcal{U} &= I(c; s, r) = H(c|s) - H(c|s, r) \\ &= H(c) + H(s, r) - H(s, r, c) \end{aligned} \quad (6)$$

The index expresses the amount of information gained about confidence from knowing the actor's response, after knowing the stimuli. When the rater's metacognitive efficiency is null, $meta - \mathcal{U}$ amounts to zero *bit*, setting a lower bound on efficiency. An upper-bound on $meta - \mathcal{U}$ is placed by the variable with the lowest number of alternatives in the communication system. Decomposition of $meta - \mathcal{U}$ into entropy terms can explicate the logic of this measure. In the first decomposition, $meta - \mathcal{U}$ reflects the degree to which the conditional uncertainty of confidence after knowing the stimulus is reduced after observing the actor's response. In the second decomposition, we note that $meta - \mathcal{U}$ gauges the degree to which uncertainty in the confidence and stimulus-response variables $H(c) + H(s, r)$ is reduced after observing the actor's response $H(s, r, c)$. This measure corrects for type 1 performance by taking the difference between type 1 uncertainty $H(s, r)$ and the type 1 and type 2 uncertainty $H(s, r, c)$, in a similar way to $meta - d' - d'$.

Importantly, it can be shown that $meta - \mathcal{U}$ equals $meta - \mathcal{I}'$ due to the symmetry of mutual information:

$$\begin{aligned} meta - \mathcal{U} &= meta - \mathcal{I}' \\ I(c; s, r) &= I(r; c, s) \end{aligned} \quad (7)$$

In the case of a two-choice SDT task, $meta - \mathcal{U}$ and $meta - \mathcal{I}'$ are equal to Dayan's $meta - \mathcal{I}$.

The other two measures advanced by Fitousi [4], $meta - \mathcal{KL}$, and $meta - \mathcal{J}$, are based on the Kullback-Leibler (KL) [50] and Jeffrey's [51] divergences. These divergences are applied to the normalized confidence-accuracy distributions $P[(confidence = y)|Correct]$ and $Q[(confidence = y)|Incorrect]$. It should be noted though that, in principle, these divergence measures can be derived for any required partition of the stimuli-response joint variable, irrespective of accuracy. Accuracy is defined by labels (i.e., "correct", "incorrect") placed by the experimenter on the actor's responses. But information-theoretic quantities are agnostic to such labels. A rater can be metacognitively efficient even when she flips the "correct" and "incorrect" labels on the actor's responses.

The new measure proposed by Fitousi, $meta - \mathcal{KL}$, is based on a KL divergence $D_{KL}(Q \parallel P)$ that measures the degree of separation between correct and incorrect accuracy-confidence distributions:

$$meta - \mathcal{KL} = D_{KL}(P \parallel Q) = \sum P(c = y|Correct) \log \left(\frac{P(c = y|Correct)}{Q(c = y|Incorrect)} \right) \quad (8)$$

It amounts to 0 *bits* when the rater cannot distinguish between correct and incorrect responses of the actor, and increases as the rater becomes more metacognitively efficient. An upper bound is met when the two distributions are at their maximal separability. An analogous expression of $meta - \mathcal{KL}$ gauges the difference between the entropy in the joint correct and incorrect distribution and the entropy in the correct distribution:

$$\begin{aligned}
meta - \mathcal{KL} &= D_{KL}(P \parallel Q) \\
&= H(P, Q) - H(P) \\
&= H(c = y|Correct, c = y|Incorrect) - H(c = y|Correct)
\end{aligned} \tag{9}$$

Metacognitive efficiency is null when uncertainty in the joint correct-incorrect confidence-accuracy distribution is equal to the uncertainty in the correct confidence-accuracy distribution. Metacognitive efficiency increases as the uncertainty in the correct distribution decreases. It reaches an upper bound when the latter is 0 *bits*, namely, the rater eliminates all uncertainty in the actor's correct (incorrect) responses to the stimuli.

The $meta - \mathcal{KL}$ measure does not satisfy the triangle inequality, and is also not symmetric in the two distributions, since $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$. In this sense, $meta - \mathcal{KL}$ does not qualify as a metric [52]. To address these constraints, Fitousi [4] has also proposed $meta - \mathcal{J}$ based on the Jeffrey's divergence [51,53], which offers a symmetric measure of divergence:

$$\begin{aligned}
meta - \mathcal{J} &= J(P, Q) = D_{KL}(P \parallel Q) + D_{KL}(Q \parallel P) \\
&= \sum P(c = y|Correct) \log \left(\frac{P(c = y|Correct)}{Q(c = y|Incorrect)} \right) + \\
&\quad \sum Q(c = y|Incorrect) \log \left(\frac{Q(c = y|Incorrect)}{P(c = y|Correct)} \right)
\end{aligned} \tag{10}$$

This measure functions in a similar way to $meta - \mathcal{KL}$. It equals 0 *bits* when the rater cannot distinguish between the correct and incorrect accuracy-confidence distributions, and increases as the rater becomes more efficient. But unlike $meta - \mathcal{KL}$, it takes into consideration the two possible KL divergences, $D_{KL}(P \parallel Q)$ and $D_{KL}(Q \parallel P)$ between the correct and incorrect confidence-accuracy distributions. Another way of expressing the same measure:

$$\begin{aligned}
meta - \mathcal{J} &= J(P, Q) \\
&= 2 \times H(P, Q) - H(P) - H(Q) \\
&= 2 \times H(c = y|Correct, c = y|Incorrect) - H(c = y|Correct) - H(c = y|Incorrect)
\end{aligned} \tag{11}$$

suggests that metacognitive efficiency gauges the degree to which the rater reduces uncertainty in the actor's response by eliminating uncertainty in the correct accuracy-confidence distribution, and/or in the incorrect accuracy-confidence distribution. The $meta - \mathcal{J}$ takes into consideration two sources of uncertainty, weighting them as two independent contributions to metacognitive efficiency. This is in contrast to the $meta - \mathcal{KL}$, which gives all the weight to the uncertainty in the correct accuracy-confidence distribution.

To validate these novel measures, Fitousi [4] has performed simulations with the $meta - d'$ model [1,31]. He varied the values of d' and $meta - d'$ incrementally in small steps, and in an orthogonal fashion and generated the corresponding confidence-accuracy distribution for each pair of values. He then computed the information-theoretic measures $meta - \mathcal{U}$, $meta - \mathcal{KL}$, and $meta - \mathcal{J}$ for each confidence-accuracy distribution, and plotted those values as a heatmap, with d' in the abscissa and $meta - d'$ in the ordinate. Figure 3 presents the results of these simulations. As can be noted, all three heatmaps reveal one of the most important properties of an efficiency measures, namely that the rater's sensitivity is corrected for type 1 performance. To see why, note that the highest values of metacognitive efficiency are achieved when the actor's sensitivity (d') is low and the rater's sensitivity ($meta - d'$) is high. In all three heatmaps, as the actor's sensitivity increases, there is less and less room for the rater to accomplish high metacognitive efficiency. Also notable are the differences between the mutual information measure ($meta - \mathcal{U}$), and the divergence measures ($meta - \mathcal{KL}$, $meta - \mathcal{J}$) with respect to the maximal values they can take, and their overall dynamics.

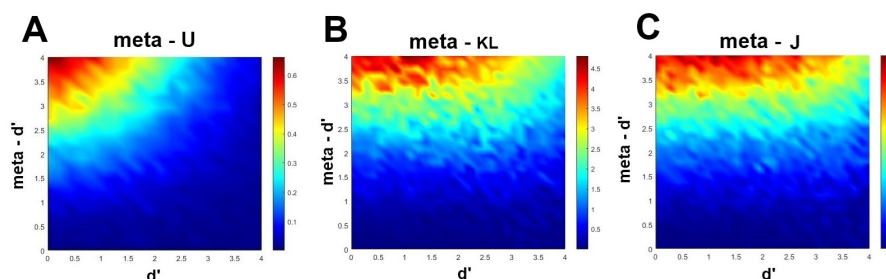


Figure 3. Theoretical validation of the information measures of metacognitive efficiency using the $meta - d'$ model [1,31]. Data are generated by increasing incrementally and in an orthogonal fashion the values of parameters d' and $meta - d'$, and then computing the corresponding values of $meta - \mathcal{U}$, $meta - \mathcal{KL}$, and $meta - \mathcal{J}$. The resulting values are presented in a heatmap against the values of d' and $meta - d'$. (A) $meta - \mathcal{U}$. (B) $meta - \mathcal{KL}$. (C) $meta - \mathcal{J}$.

The information theoretic measures have many valuable characteristics. These are discussed in detail in the papers by Dayan [3] and Fitousi [4]. Here I give a brief summary of the most important features. They are: (a) relative measures, (b) accommodate type 1 performance, (c) bias-free, (d) do not require any parametric assumptions, (e) scale with the number of confidence rating points, (f) do not require the assumption of ratio or order of confidence ratings, (g) easy to compute, and (h) offer definite lower and upper bounds on performance.

One important question that remains to be address is whether these measures of metacognitive efficiency converge on the same psychological construct. Previous efforts by Dayan [3] and Fitousi [4] have emulated various simulations with theoretical models to show that the measures converge. But there is still the need to assess these measures' convergence with real data sets. If these different indexes capture the same psychological construct, they should correlate in real empirical data. This hypothesis is subjected to scrutiny in the next section.

3. Empirical Validation

3.1. The Face-Matching Task

Here I provide an empirical validation of the information-theoretic measures of metacognitive efficiency. To do so, I take advantage of two data sets collected in a recent face-matching study in my lab. The face-matching task [54–58] has become an important tool in assessing sundry aspects of face recognition. In this task participants are presented with two face images and are asked to decide whether these belong to the same identity, or alternatively to two different identities. The images are taken under various lighting conditions, with different cameras, and at different times. It turns out that with unfamiliar facial identities, this task poses serious challenges for both humans [21,54–58] and machines [59]. This is mainly because under variable ambient conditions, images of the same personal identity can look completely different; whereas images of different identities may appear similar.

Figure 4 presents examples of pairs of face images from the Glasgow Face Matching Task [GFMT 54], and from the Kent Face Matching Task [KFMT 60], two of the most popular standard tests, which were also deployed in the present investigation. Extensive research on the face-matching task [61,62] has greatly enhanced our understanding of the underlying mechanisms that govern this task. A recent paper by Fitousi [21] has introduced an unequal-variance signal detection model, one that can account for many aspects of this task, including type 1 confidence judgments, similarity ratings, and accuracy.

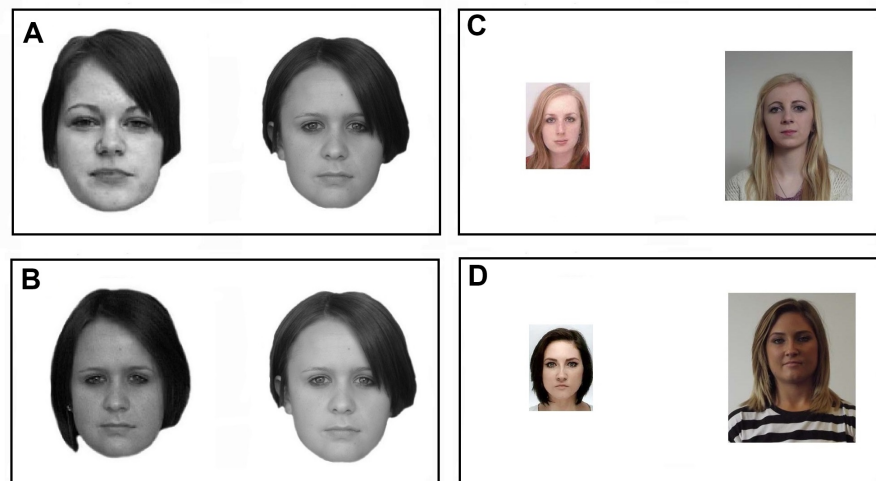


Figure 4. Examples of face-pairs from the Glasgow Face Matching Task (GFMT, panels A-B) and from the Kent Face Matching Task (KFMT, panels C-D). Panels A and C depict the same identity (i.e., “match”), whereas panels B and D depict different identities (i.e., “mismatch”).

However, despite extensive research on the face-matching task [55,63–65], only little is known about the underlying metacognitive capabilities with which participants perform this task. The application of the *meta* – d' indexes, along with the novel information-theoretic measures can shed new light on this issue. But most importantly, the joint application of these measures is needed to garner strong evidence for their convergence.

3.2. The Data Sets

The data sets were collected in my lab. In data set 1, seventy participants engaged in a face-matching task with the long version of the GFMT [54]. In data set 2, seventy new participants were tested with the KFMT [60]. There were 168 images in the GFMT and 40 images in the KFMT. More details about these tests can be found here [21]. The general method in both experiments was identical. Participants were sitting in front of a computer screen. On each trial they were presented with a randomly selected face-pair, and were asked to decide as accurately and as speedily as they can, whether the pair constituted a “match” or a “mismatch” by pressing one of two computer keys (type 1 task). Immediately after responding, the images were removed from the screen, and a four-point confidence scale appeared on the screen with 1=“not confident” and 4=“highly confident.” In data set 1 participants completed two blocks, which amounted to a total of 336 trials. In data set 2 participants completed four blocks, which resulted in a total of 160 trials.

3.3. Results

The face-matching task administrated in data set 1 was relatively easier ($d' = 2.51$, $sd = 0.65$) than the one in data set 2 ($d' = 1.01$, $sd = 0.41$). However, the results from both data sets converged on the conclusion that, on average, participants did exhibit meta-knowledge about their face-matching performance. The mean of the *meta* – $d' - d'$ measure amounted to 0.86 ($sd = 1.31$) in data set 1, and 1.09 ($sd = 0.97$) in data set 2. Recall that for *meta* – $d' - d'$ a value greater than zero entails hypersensitivity. Similarly, the mean of the *M* – *ratio* measure was equal to 1.44 ($sd = 0.66$) in data set 1, and 2.83 ($sd = 3.25$) in data set 2. Recall that for *M* – *ratio* a value greater than 1 entails hypersensitivity. One-sided student-tests supported these conclusions, with all *t*-statistics being significant at the 0.05 level. The novel information-theoretic measures also supported the conclusion that participants in both data sets acquired meta-knowledge about their decisions. The average *meta* – $\mathcal{U}$ values were 0.22 bits ($sd = 0.17$) and 0.37 bits ($sd = 0.21$) in the two data sets, respectively. Recall that in the two-choice task *meta* – $\mathcal{U}$ is quantitatively comparable to Dayan’s [3] *meta* – $\mathcal{I}$. The average *meta* – $\mathcal{KL}$ values

amounted to 1.54 bits (sd = 0.95) and 1.11 bits (sd = 0.74) in the two data sets, respectively. The average $meta - \mathcal{J}$ values were 4.15 bits (sd = 2.37) and 2.54 bits (sd = 1.74) in the two data sets, respectively.

I then computed the SDT and information-theory measures for each participant, and used linear regression on each pair of measures to quantify the *coefficient of determination* R^2 . Figures 5 and 6 present the results of these analyses, respectively, for the GFMT and KFTM data sets.

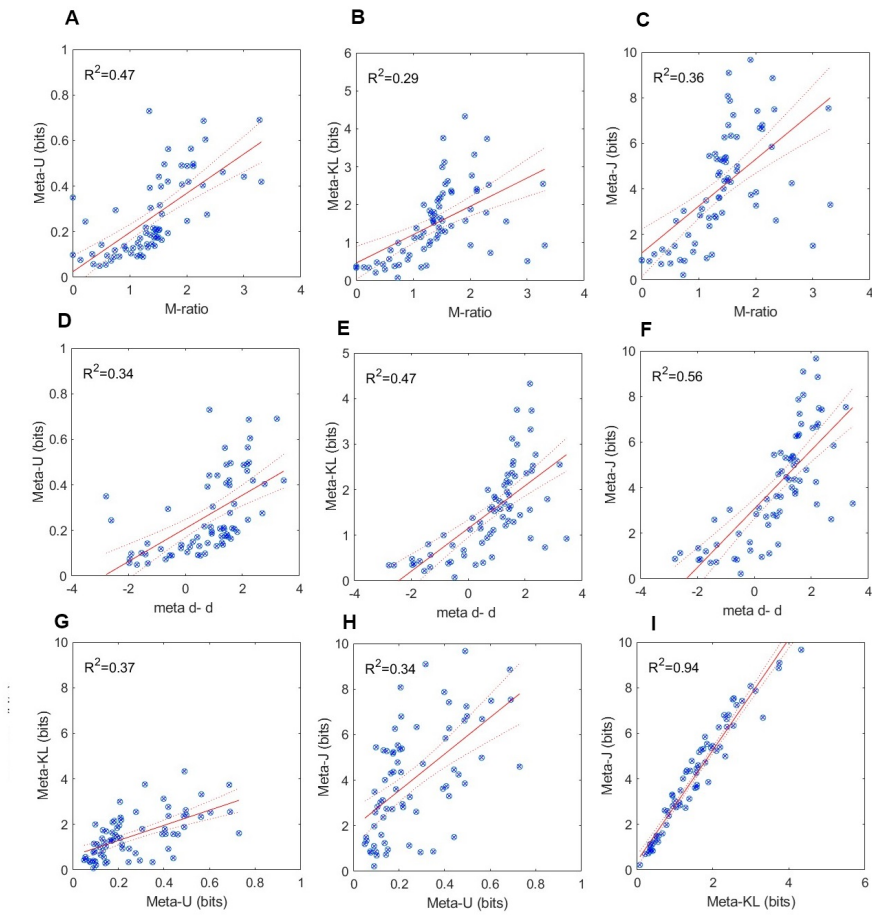


Figure 5. Results from the GFMT (Glasgow Face Matching Task): Relations between SDT and information theoretic measures of metacognitive efficiency. (A)-(C) M-ratio as a function of meta-U, meta-KL and meta - J. (D)-(F) meta $d' - d'$ as a function of meta-U, meta-KL and meta - J. (G)-(I) relations between meta-U, meta-KL and meta - J.

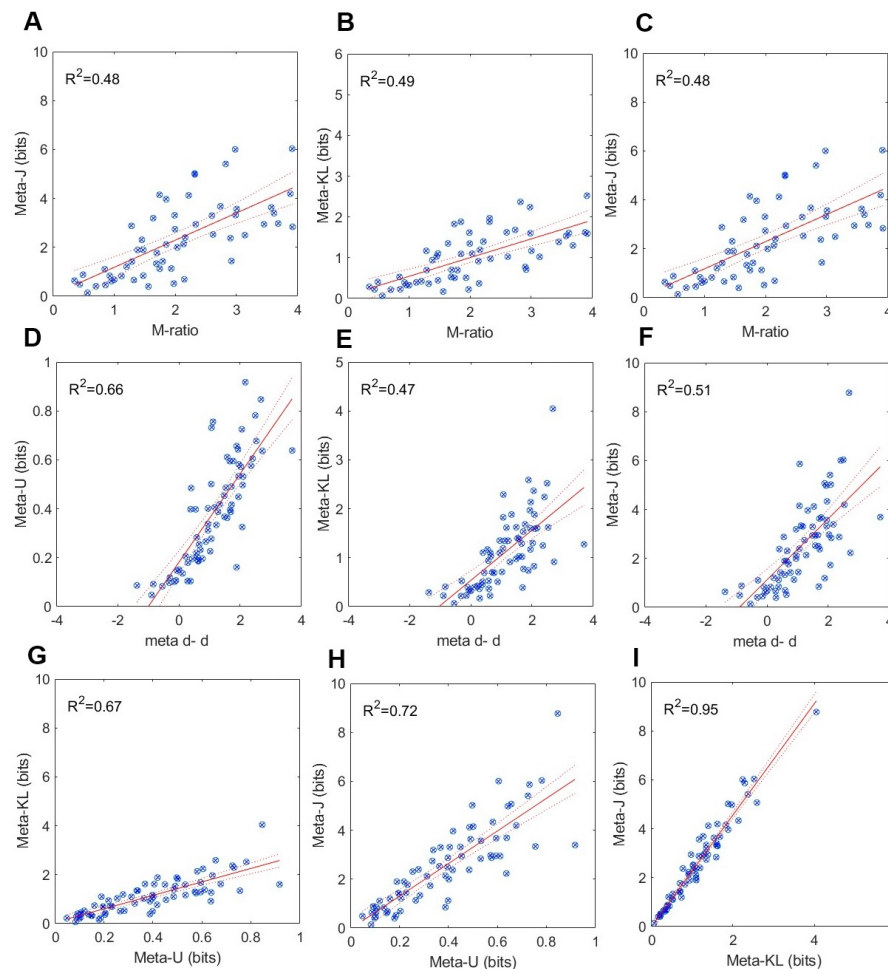


Figure 6. Results from the KFMT (Kent Face Matching Task): Relations between SDT and information theoretic measures of metacognitive efficiency. (A)-(C) M-ratio as a function of meta-U, meta-KL and meta - J. (D)-(F) meta $d' - d'$ as a function of meta-U, meta-KL and meta - J. (G)-(I) relations between meta-U, meta-KL and meta - J.

As can be noted in Figures 5 and 6, significant positive and strong relations ensued between each pair of measures in both data sets. Overall, higher R^2 obtained for the KFMT than the GFMT data. In the former, all coefficient of determination values were larger than 0.47. In the latter, all those values exceeded 0.29. The strongest relations were observed between $meta - \mathcal{KL}$ and $meta - \mathcal{J}$. This is quite expected given their mathematical affinity. Taken collectively, these results: (a) support the conclusion that participants do have meta-knowledge about their face-matching performance, and (b) confer validation support in the convergence of the information-theoretic measures on a common psychological structure – metacognitive efficiency.

4. Conclusions

I introduced various information-theoretic measures of metacognitive efficiency recently developed by Dayan [3] and Fitousi [4]. These indexes consist of: the $meta - \mathcal{I}$ [3], and the $meta - \mathcal{I}'$, $meta - \mathcal{U}$, $meta - \mathcal{KL}$, and $meta - \mathcal{J}$ [4]. These measures were originally developed within a common theoretical framework, assuming a communication system with stimuli (experimenter), actor, and rater as encoders/decoders. I provided a succinct account of each measure, along with its possible interpretations. I then applied these information indexes, along with the $M - ratio$ and $meta - d' - d'$ indexes, to data sets from the face-matching task [54]. The results strongly supported the convergence of these measures on a single psychological construct.

One source of motivation for developing information-theoretic markers of metacognitive efficiency has to do with the limitations of existing model-based measures, such as $metad' - d'$ and $M - ratio$.

These indexes are built on the assumption that they should fit exactly to a type 1 decision process. But this assumption is dubious [3]. Moreover, these measures are subjected to decisional bias [1,2], have a remaining dependency on type 1 performance [33], and on metacognitive bias [34]. Information theory measures are free of such limitations.

Another crucial point is that both model-based and model-free measures of metacognitive efficiency [66] do not accommodate changes in the granularity of type 2 confidence ratings, nor do they scale with the number of type 1 stimuli or responses. Information-theoretic measures can be computed relatively easily, without making any parametric assumptions, can be applied to tasks which incorporate multi-alternative choices, are compatible with existent SDT-based measures, and can be applied to data without labeling the responses as correct or incorrect. Moreover, unlike SDT-based measures, they can accommodate the granularity of the confidence judgments (e.g., number of rating points), and can provide definite upper- and lower-bounds on performance. All of these valuable characteristics have been previously demonstrated by Dayan [3] and Fitoussi [4].

Another reason for advancing the information-theoretic approach to metacognition has to do with the suitability of such a theory to characterize capacity limitations, information-bottlenecks, rate-distortions, and other quantities that are not readily available in current approaches (e.g. SDT). Future work on confidence and metacognition should seek to expand our theoretical understanding by building on these powerful notions.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Maniscalco, B.; Lau, H. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and cognition* **2012**, *21*, 422–430.
2. Maniscalco, B.; Lau, H. Signal detection theory analysis of type 1 and type 2 data: meta-d, response-specific meta-d, and the unequal variance SDT model. In *The Cognitive Neuroscience of Metacognition*; Springer, 2014; pp. 25–66.
3. Dayan, P. Metacognitive information theory. *Open Mind* **2023**, *7*, 392–411.
4. Schulz, L.; Rollwage, M.; Dolan, R.J.; Fleming, S.M. Dogmatism manifests in lowered information search under uncertainty. *Proceedings of the National Academy of Sciences* **2020**, *117*, 31527–31534.
5. Schulz, L.; Fleming, S.M.; Dayan, P. Metacognitive computations for information search: Confidence in control. *Psychological Review* **2023**, *130*, 604–639.
6. Peirce, C.S.; Jastrow, J. On small differences in sensation. *Memoirs of the National Academy of Sciences* **1884**, 3–11.
7. Henmon, V.A.C. The relation of the time of a judgment to its accuracy. *Psychological Review* **1911**, *18*, 186–201.
8. Lau, H. In *consciousness we trust: The cognitive neuroscience of subjective experience*; Oxford University Press, 2022.
9. Metcalfe, J.; Shimamura, A.P. *Metacognition: Knowing about knowing*; MIT press, 1994.
10. Fleming, S.M.; Dolan, R.J. The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences* **2012**, *367*, 1338–1349.
11. Fleming, S.M.; Weil, R.S.; Nagy, Z.; Dolan, R.J.; Rees, G. Relating introspective accuracy to individual differences in brain structure. *Science* **2010**, *329*, 1541–1543.
12. Fleming, S.M. Metacognition and confidence: A review and synthesis. *Annual Review of Psychology* **2024**, *75*, 241–268.
13. Meyniel, F.; Schlunegger, D.; Dehaene, S. The sense of confidence during probabilistic learning: A normative account. *PLoS computational biology* **2015**, *11*, e1004305.
14. Koriat, A.; Goldsmith, M. Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review* **1996**, *103*, 490–517.
15. Koriat, A. Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General* **1997**, *126*, 349–370.
16. Koriat, A. The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and Cognition* **2000**, *9*, 149–171.

17. Persaud, N.; McLeod, P.; Cowey, A. Post-decision wagering objectively measures awareness. *Nature neuroscience* **2007**, *10*, 257–261.
18. Van den Berg, R.; Zylberberg, A.; Kiani, R.; Shadlen, M.N.; Wolpert, D.M. Confidence is the bridge between multi-stage decisions. *Current Biology* **2016**, *26*, 3157–3168.
19. Hoven, M.; Lebreton, M.; Engelmann, J.B.; Denys, D.; Luigjes, J.; van Holst, R.J. Abnormalities of confidence in psychiatry: an overview and future perspectives. *Translational psychiatry* **2019**, *9*, 268.
20. Bahrami, B.; Olsen, K.; Latham, P.E.; Roepstorff, A.; Rees, G.; Frith, C.D. Optimally interacting minds. *Science* **2010**, *329*, 1081–1085.
21. Fitousi, D. A Signal-detection based confidence-similarity model of face-matching. *Psychological Review* **2024**, *131*, 625–663.
22. Garner, W.R.; Hake, H.W.; Eriksen, C.W. Operationism and the concept of perception. *Psychological Review* **1956**, *63*, 149–159.
23. Clarke, F.R.; Birdsall, T.G.; Tanner Jr, W.P. Two types of ROC curves and definitions of parameters. *The Journal of the Acoustical Society of America* **1959**, *31*, 629–630.
24. Galvin, S.J.; Podd, J.V.; Drga, V.; Whitmore, J. Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review* **2003**, *10*, 843–876.
25. Fleming, S.M.; Daw, N.D. Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review* **2017**, *124*, 91–114.
26. Green, D.M.; Swets, J.A. *Signal detection theory and psychophysics*; Vol. 1, Wiley New York, 1966.
27. Macmillan, N.A.; Creelman, C.D. *Detection theory: A user's guide*; Psychology press, 2004.
28. Nelson, T.O. A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin* **1984**, *95*, 109–133.
29. Fleming, S.M. HMeta-d: hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of consciousness* **2017**, *2017*, nix007.
30. Kunimoto, C.; Miller, J.; Pashler, H. Confidence and accuracy of near-threshold discrimination responses. *Consciousness and cognition* **2001**, *10*, 294–340.
31. Maniscalco, B.; Peters, M.A.; Lau, H. Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Attention, Perception, & Psychophysics* **2016**, *78*, 923–937.
32. Shekhar, M.; Rahnev, D. Sources of metacognitive inefficiency. *Trends in Cognitive Sciences* **2021**, *25*, 12–23.
33. Guggenmos, M. Measuring metacognitive performance: type 1 performance dependence and test-retest reliability. *Neuroscience of consciousness* **2021**, *2021*, niab040.
34. Xue, K.; Shekhar, M.; Rahnev, D. Examining the robustness of the relationship between metacognitive efficiency and metacognitive bias. *Consciousness and Cognition* **2021**, *95*, 103196.
35. Luce, R.D. Whatever happened to information theory in psychology? *Review of general psychology* **2003**, *7*, 183–188.
36. Knill, D.C.; Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences* **2004**, *27*, 712–719.
37. Feldman, J. Information-theoretic signal detection theory. *Psychological Review* **2021**, *128*, 976–987.
38. Killeen, P.; Taylor, T. Bits of the ROC: Signal detection as information transmission. *Unpublished manuscript* **2001**.
39. Fitousi, D. Mutual information, perceptual independence, and holistic face perception. *Attention, Perception, & Psychophysics* **2013**, *75*, 983–1000.
40. Fitousi, D.; Algom, D. The quest for psychological symmetry through figural goodness, randomness, and complexity: A selective review. *i-Perception* **2024**, *15*, 1–24.
41. Fitousi, D. Quantifying Entropy in Response Times (RT) Distributions Using the Cumulative Residual Entropy (CRE) Function. *Entropy* **2023**, *25*, 1239.
42. Friston, K. The free-energy principle: a unified brain theory? *Nature reviews neuroscience* **2010**, *11*, 127–138.
43. Norwich, K.H. *Information, sensation, and perception*; Academic Press San Diego, 1993.
44. Hirsh, J.B.; Mar, R.A.; Peterson, J.B. Psychological entropy: A framework for understanding uncertainty-related anxiety. *Psychological Review* **2012**, *119*, 304–320.
45. Shannon, C. A mathematical theory of communication. *The Bell system technical journal* **1948**, *27*, 379–423.
46. Shannon, C.; Weaver, W. *The Mathematical Theory of Communication*, (first published in 1949). *Urbana University of Illinois Press* **1949**.
47. Friston, K. The free-energy principle: a rough guide to the brain? *Trends in cognitive sciences* **2009**, *13*, 293–301.

48. Pouget, A.; Drugowitsch, J.; Kepecs, A. Confidence and certainty: distinct probabilistic quantities for different goals. *Nature neuroscience* **2016**, *19*, 366–374.
49. Li, H.H.; Ma, W.J. Confidence reports in decision-making with multiple alternatives violate the Bayesian confidence hypothesis. *Nature communications* **2020**, *11*, 2004.
50. Kullback, S. *Information theory and statistics*; Courier Corporation, 1959.
51. Jeffrey, R.C. *Probability and the Art of Judgment*; Cambridge University Press, 1992.
52. Cover, T.; Thomas, J. *Elements of information theory*; John Wiley & Sons, 1991.
53. Pearl, J. Jeffrey's rule, passage of experience, and neo-Bayesianism. In *Knowledge representation and defeasible reasoning*; Springer, 1990; pp. 245–265.
54. Burton, A.M.; White, D.; McNeill, A. The Glasgow face matching test. *Behavior research methods* **2010**, *42*, 286–291.
55. White, D.; Towler, A.; Kemp, R. Understanding professional expertise in unfamiliar face matching. *Forensic face matching: Research and practice* **2021**, pp. 62–88.
56. Megreya, A.M.; Burton, A.M. Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition* **2006**, *34*, 865–876.
57. Megreya, A.M.; Burton, A.M. Hits and false positives in face matching: A familiarity-based dissociation. *Perception & Psychophysics* **2007**, *69*, 1175–1184.
58. Kemp, R.; Towell, N.; Pike, G. When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* **1997**, *11*, 211–222.
59. O'Toole, A.J.; An, X.; Dunlop, J.; Natu, V.; Phillips, P.J. Comparing face recognition algorithms to humans on challenging tasks. *ACM Transactions on Applied Perception (TAP)* **2012**, *9*, 1–13.
60. Fysh, M.C.; Bindemann, M. The Kent face matching test. *British Journal of Psychology* **2018**, *109*, 219–231.
61. Bate, S.; Mestry, N.; Portch, E. Individual differences between observers in face matching. *Forensic Face Matching: Research and Practice* **2021**, pp. 117–145.
62. Towler, A.; White, D.; Kemp, R.I. Evaluating training methods for facial image comparison: The face shape strategy does not work. *Perception* **2014**, *43*, 214–218.
63. Jenkins, R.; White, D.; Van Montfort, X.; Burton, A.M. Variability in photos of the same face. *Cognition* **2011**, *121*, 313–323.
64. Bindemann, M.; Attard, J.; Leach, A.; Johnston, R.A. The effect of image pixelation on unfamiliar-face matching. *Applied Cognitive Psychology* **2013**, *27*, 707–717.
65. Megreya, A.M.; Bindemann, M. Feature instructions improve face-matching accuracy. *PloS one* **2018**, *13*, e0193455.
66. Shekhar, M.; Rahnev, D. The nature of metacognitive inefficiency in perceptual decision making. *Psychological Review* **2021**, *128*, 45–70.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.