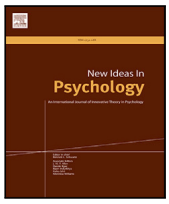
New Ideas in Psychology 82 (2026) 101249

Contents lists available at ScienceDirect

New Ideas in Psychology

journal homepage: www.elsevier.com/locate/newideapsych



Accumulated evidence as an additive performance measure based on confidence ratings☆

Sascha Meyen *, Lin Lin, Carina Schrenk, Volker H. Franz

Department of Computer Science, University of Tübingen, Germany

ARTICLE INFO

Dataset link: osf.io/krw53

Keywords:
Evidence accumulation
Drift diffusion models
Sequential probability ratio test

ABSTRACT

Perceptual evidence accumulation is a gradual process that is typically studied using drift diffusion models. A disadvantage of those models is that they treat accumulated evidence only in an indirect way. Here, we suggest to measure accumulated evidence more directly by utilizing participants' confidence ratings together with their decisions. The resulting measure has the key advantage of being additive: For example, evidence accumulated during a first stimulus presentation interval $E_{[0,a]}$ adds up with evidence accumulated during a second interval $E_{[a,b]}$ to the total evidence accumulated during the whole presentation $E_{[0,a]} + E_{[a,b]} = E_{[0,b]}$. Because it is experimentally only possible to measure evidence after the first interval ($E_{[0,a]}$) or after the whole presentation ($E_{[0,b]}$), but not selectively for the second interval ($E_{[a,b]}$), we can exploit additivity to calculated this value $E_{[a,b]} = E_{[0,b]} - E_{[0,a]}$. In two experiments, we demonstrate the strengths of our approach. In Experiment 1, we provide direct support for previous findings suggesting a deceleration in evidence accumulation ($E_{[0,a]} > E_{[a,b]}$). Moreover, we show an increased evidence accumulation rate for causal rather than anti-causal dot-motion stimuli. In Experiment 2, we quantify inter-dependencies of evidence accumulation in later stages ($E_{[a,b]}$) on that of earlier stages ($E_{[0,a]}$). Limitations of our approach arise from its reliance on confidence ratings. But when properly implemented, it allows the generation of more specific hypotheses than other approaches: Instead of only testing whether evidence accumulation changes, our measure supports precise point predictions and interpretations about the particular amount of accumulated evidence.

With the introduction of drift diffusion models (DDMs, Ratcliff, 1978; Ratcliff & McKoon, 2008, see Stone, 1960, and Laming, 1968, for precursors), perceptual processing has been understood as a gradual accumulation process (for an introduction see Forstmann et al., 2016; Schwarz, 2022). According to this view, the visual system does not instantaneously recognize objects but rather incrementally collects bits of evidence until recognition is achieved. This view has produced a host of more elaborate DDMs (Ratcliff & Rouder, 2000; Usher & McClelland, 2001).

DDMs treat the internally accumulated evidence as latent. Instead of measuring this evidence directly, these models fit the parameters governing the evidence accumulation process (drift rate, starting point, boundaries, etc.) by matching them to behavioral measures (responses, response times, confidence, etc.). For this latent parameter estimation, increasingly sophisticated methods have been developed (Richter et al., 2023; Shinn et al., 2020). In trying to make the accumulated evidence more tangible, many studies relate it to neuroimaging data (Fleming et al., 2018; Gold et al., 2008; Gold & Shadlen, 2007; Kiani & Shadlen, 2009; Newsome et al., 1989; Philiastides et al., 2006; Purcell et al., 2010; Stockart et al., 2024) or pupil measurements (Meyniel, 2020).

As a complementary approach, we suggest measuring accumulated evidence explicitly from participants' behavioral responses. This approach is based on sequential probability ratio tests (SPRTs; Griffith et al., 2021; Kullback, 1959; Wald, 1947; Wald & Wolfowitz, 1948), whose principles also underlie DDMs (Bitzer et al., 2014; Dayan & Daw, 2008; Gold & Shadlen, 2007; Moran, 2015). In SPRTs, the logarithm of the ratio of probabilities corresponds to the evidence of individual pieces of information that can be summed up to yield the total amount of evidence. This mathematical relationship is frequently used in modeling sequential accumulation of information (Dayan & Daw, 2008; Fleming et al., 2018; Kiani & Shadlen, 2009; Lange et al., 2021). At its core, it requires knowledge about the probabilities of participants' decisions in each trial. As a stand-in for these probabilities in SPRT, we use confidence ratings—after calibrating them—to compute the signed logarithmic odds to measure the amount of accumulated evidence.

☆ This article is part of a Special issue entitled: 'Probabilistic cognition' published in New Ideas in Psychology.

* Correspondence to: University of Tübingen, Sand 6, 72076 Tübingen, Germany.
E-mail address: sascha.meyen@uni-tuebingen.de (S. Meyen).

https://doi.org/10.1016/j.newideapsych.2026.101249
Received 30 April 2025; Received in revised form 5 November 2025; Accepted 19 February 2026
Available online 23 February 2026

How participants give these confidence ratings has been recently studied with great interest in perceptual decision tasks (Fleming, 2024; Guggenmos, 2022; Peters, 2022; Pleskac & Busemeyer, 2010; Rahnev et al., 2022; Rausch et al., 2018; Sánchez-Fuenzalida et al., 2025). Intriguingly, many of these studies show that confidence ratings are affected by a variety of factors other than the sensory information on which perceptual decisions are based. For example, additional perceptual processing after the decision is made (Hellmann et al., 2023, 2024; Moran et al., 2015; Navajas et al., 2016; Pleskac & Busemeyer, 2010), stimulus visibility (Hellmann et al., 2023, 2024; Rausch et al., 2018), response time (Hellmann et al., 2024; Kiani et al., 2014), and confidence response time (Moran et al., 2015) influence confidence. Nevertheless, confidence ratings are somewhat informative about the uncertainty in the internal processing of the stimulus (Kepecs & Mainen, 2012; Kiani & Shadlen, 2009; Peters, 2022) and are also tied to neurological measures of evidence (Dou et al., 2024; Gherman & Philiastides, 2015; Hebart et al., 2014). Thus, despite the partial misalignment of these confidence ratings with sensory information, we use them to compute a measure of evidence based on participants' behavioral responses.

The main advantage of this approach lies in its additivity. When defined in this way, evidence accumulated in two intervals of stimulus presentation sums up: If $E_{[0,a]}$ is the evidence that is accumulated while a stimulus is presented for an interval between 0 ms (onset) and a ms, and $E_{[a,b]}$ is the evidence that is subsequently accumulated in the interval between a ms and b ms, then the total evidence accumulated throughout the whole presentation duration from 0 ms to b ms is the sum $E_{[0,a]} + E_{[a,b]} = E_{[0,b]}$. This formulation allows access to the amount of evidence participants accumulate during the second interval $E_{[a,b]}$, which is not directly observable otherwise. Due to additivity, this evidence simply calculates as $E_{[a,b]} = E_{[0,b]} - E_{[0,a]}$. We can then compare $E_{[0,a]}$ to $E_{[a,b]}$ to make inferences about the trajectory of the evidence accumulation process.

In contrast to this approach, DDMs do not require trial-wise confidence ratings but instead—like their non-sequential siblings, signal detection theory models (Green & Swets, 1988; Griffith et al., 2021)—assume multiple normally distributed decision variables to sequentially add up for the total amount of evidence (Bitzer et al., 2014; Ratcliff, 1978; Ratcliff et al., 2016). This is in line with the SPRTs framework because, with underlying normal distributions, decision variables are linearly related to the logarithmic odds and can therefore be added up without loss of information. Thus, the normal distribution assumption grants DDMs their simplicity. But although the assumption of underlying normal distributions is often embraced in visual processing and justified by the central limit theorem (Schwarz, 2022; Softky & Koch, 1993; Usher & McClelland, 2001), work on Lévy-flight models has recently suggested that internal noise distributions can have heavier-than-normal tails (Rasanan et al., 2023; Voss et al., 2019; Wieschen et al., 2020). In these cases, interpretations about the evidence accumulation trajectories could be biased as we will demonstrate below. Instead of relying on distributional assumptions, we will rely on participants' reported confidence ratings to derive an explicit measure of the accumulated evidence per trial.

We first summarize the mathematical principles based on SPRTs that underlie our approach of measuring evidence based on confidence ratings. We demonstrate its appeal in theoretical examples and apply this measure in two experiments. In Experiment 1, we validate it and, additionally, implement a manipulation that allows presenting stimuli in a causal (stimulus motion played forward) versus anti-causal condition (motion played backward) showing an advantage in evidence accumulation for the causal motion. In Experiment 2, we investigate how later stages of evidence accumulation depend on earlier stages. In both experiments, measuring additive evidence allows to specify and test hypotheses more precisely than with other performance measures.

1. Evidence accumulation based on calibrated confidence ratings

We consider a setting in which stimuli belong to either of two categories, Y. In our experiments, this corresponds to the true motion direction of dot stimuli. The concrete values are denoted by lowercase variables y, which here can take -1 (leftward motion) or $+1$ (rightward motion).

Assumption 1 (*Binary Label*). $Y \in \{-1, +1\}$

We assume both categories to be presented with equal probabilities.

Assumption 2 (*Equal Probability*). $P(Y = +1) = P(Y = -1) = 0.5$

Participants respond by giving a decision $\hat{Y}$, a binary decision about what the true stimulus category is believed to be, accompanied by a confidence rating C in each trial.

Assumption 3 (*Binary Decision*). $\hat{Y} \in \{-1, +1\}$

For the theoretical derivation, we assume the confidence ratings to correspond to the posterior probability of a decision being correct (Boundy-Singer et al., 2023; Kepecs & Mainen, 2012; Koriat, 2011; Meyniel et al., 2015; Pouget et al., 2016). Although participants have access to internal uncertainties when integrating information internally (Balsdon et al., 2020; de Gardelle & Summerfield, 2011; Fetsch et al., 2011; Hausmann & Läge, 2008; Ma et al., 2006; Miyoshi et al., 2025; Peters et al., 2017; Whiteley & Sahani, 2008), assuming they can report perfectly calibrated confidence ratings is not realistic (Gigerenzer et al., 1991; Phillips & Edwards, 1966; Rahnev & Denison, 2018; Zhang & Maloney, 2012). Thus, we later calibrate confidence ratings to ensure this assumption holds roughly. For now, we make the simplifying, strong assumption that we can obtain such calibrated confidence ratings for the purpose of the theoretical derivations.

Assumption 4 (*Calibrated Confidence*). $c = P(Y = \hat{y} \mid C = c, \hat{Y} = \hat{y})$.

In our experiments, we present participants with stimuli for either a short or a long duration. The short duration is represented by the interval $[0, a]$ indicating that the stimulus is presented from onset (0) to a time point a. Based on this short presentation, a participant makes a decision $\hat{Y}_{[0,a]}$ and gives a confidence rating $C_{[0,a]}$ about that decision. Similarly, the long duration represented by the interval $[0, b]$ (with $a < b$) gives rise to a decision $\hat{Y}_{[0,b]}$ and confidence rating $C_{[0,b]}$. We are interested in the evidence that is accumulated exclusively in the second interval of the long stimulus presentation, $[a, b]$, which is not directly observable. However, using a definition of evidence that allows for a simple algebraic decomposition, we will extract this evidence which corresponds to the unobservable variables $\hat{Y}_{[a,b]}$ and $C_{[a,b]}$ reflecting the evidence participants collected in the second interval of stimulus presentation.

Disentangling the evidence that is accumulated in early and late stages is theoretically interesting because we know that there are dependencies of the late on the early stages. For example with more complex stimuli, the reverse hierarchy theory suggests initial processing of low spatial frequency features to guide subsequent processing of high spatial frequency features (Hochstein & Ahissar, 2002; Kauffmann et al., 2015; Rao & Ballard, 1999). More generally, predictive coding theories assume that early visual processing changes, through feedback, how subsequent information is processed (Rao & Ballard, 1999). Similar feedback processes have been empirically observed (Murphy et al., 2015). Quantifying these dependencies will allow for more precise hypothesis generation. While some experiments investigate this question by querying participants for responses after the first stimulus interval and then again after the second (Fleming et al., 2018; Rollwage et al., 2020), this disrupts the continuous perceptual process.

Finally, we assume that the evidence from each of the two intervals yields independent information about the stimulus category Y. This

does not contradict the idea of dependencies between the amounts of information gathered in the two intervals: A high amount of information in the first interval may allow for better information gathering in the second. However, the information about Y from the two intervals must compound and not be redundant.

Assumption 5 (*Non-Redundant Information*).

$$P\left(Y \mid C_{[0,a]}, \hat{Y}_{[0,a]}, C_{[a,b]}, \hat{Y}_{[a,b]}\right) = \frac{P(Y|C_{[0,a]},\hat{Y}_{[0,a]})P(Y|C_{[a,b]},\hat{Y}_{[a,b]})}{\sum_{y\in\pm1} P(y|C_{[0,a]},\hat{Y}_{[0,a]})P(y|C_{[a,b]},\hat{Y}_{[a,b]})}$$

Under these assumptions, evidence can be measured in an additive way.

1.1. Additive evidence in a single trial

Consider a hypothetical experiment following Assumptions 1–5 in which a participant discriminates the direction of dot motion. Suppose rightward dot motion is shown in one trial ($Y = +1$). After a short presentation duration a, the researcher were able to freeze time and ask the participant for their responses. Say, the participant veridically perceived the dots moving right, $\hat{Y}_{[0,a]} = +1$, with a confidence rating of $C_{[0,a]} = 75\%$. The researcher now unfroze time and let the stimulus presentation continue and the participant perceived it as if they were never interrupted until time point b. Then, the participant gave their final response as having perceived rightward motion, $\hat{Y}_{[0,b]} = +1$, with a now higher confidence rating $C_{[0,b]} = 90\%$. From this, the researcher could infer that the second interval provided additional evidence equivalent to an independent observation supporting rightward motion, $\hat{Y}_{[a,b]} = +1$, also with confidence rating $C_{[a,b]} = 75\%$. This is so because two independent guesses for the same label with the confidence rating 75% yield a combined confidence of 90% ($\frac{.75\cdot.75}{.75\cdot.75+.25\cdot.25} = .9$).

$$\begin{aligned} &\hat{Y}_{[0,a]} = +1, \quad C_{[0,a]} = .75 \\ &\hat{Y}_{[a,b]} = +1, \quad C_{[a,b]} = .75 \\ \rightarrow\ &\hat{Y}_{[0,b]} = +1, \quad C_{[0,b]} = .90 \end{aligned}$$

To allow inferring the evidence accumulated in the second interval of stimulus presentation more conveniently and with advantageous statistical properties, we will translate participants' responses into an additive measure of evidence. Based on the principles of SPRTs, we define evidence as the signed logarithmic odds of (calibrated) confidence ratings.

Definition 1 (*Evidence*).

$$E = Y \cdot \hat{Y} \cdot \text{logit}(C) \quad with \quad \text{logit}(C) = \log\left(\frac{C}{1-C}\right)$$

Under this definition, evidence E is positive when a decision is correct ($Y \cdot \hat{Y} = +1$ either because both, true label and decision, are $+1$ or both are -1). Otherwise, evidence is negative corresponding to an incorrect decision.

When transforming responses into evidence in this way, combining independent pieces of evidence becomes additive: Evidence from presenting a stimulus in a first interval $E_{[0,a]}$ adds to the evidence from presenting independent information in a second interval $E_{[a,b]}$ for a total of $E_{[0,a]} + E_{[a,b]} = E_{[0,b]}$. In the literature, this relationship is often written out as

$$\log\left(\frac{p_{\text{combined}}}{1-p_{\text{combined}}}\right) = \log\left(\frac{p_1}{1-p_1}\right) + \log\left(\frac{p_2}{1-p_2}\right)$$

or similar (Dayan & Daw, 2008; Fleming et al., 2018; Jaynes, 2003; Lange et al., 2021) where p_1, p_2, and p_{combined} often represent prior, likelihood, and posterior beliefs, respectively.

Proposition 2 (*Evidence Additivity*). *Under Assumptions 1–5, responses* $\left(\hat{Y}_{[0,a]}, C_{[0,a]}\right)$ *and* $\left(\hat{Y}_{[a,b]}, C_{[a,b]}\right)$ *correspond to evidence values*

$E_{[0,a]} = Y \cdot \hat{Y}_{[0,a]} \cdot \text{logit}(C_{[0,a]})$ *and*

$E_{[a,b]} = Y \cdot \hat{Y}_{[a,b]} \cdot \text{logit}(C_{[a,b]})$.

The combined evidence is

$$E_{[0,b]} = E_{[0,a]} + E_{[a,b]}.$$

where $E_{[0,b]} = Y \cdot \hat{Y}_{[0,b]} \cdot \text{logit}(C_{[0,b]})$ *corresponds to the combined response* $\left(\hat{Y}_{[0,b]}, C_{[0,b]}\right)$ *with*

$$\hat{Y}_{[0,b]} = \text{sign}\left(\hat{Y}_{[0,a]} \cdot \text{logit}\left(C_{[0,a]}\right) + \hat{Y}_{[a,b]} \cdot \text{logit}\left(C_{[a,b]}\right)\right)$$

$$C_{[0,b]} = \text{logit}^{-1}\left(\left|\hat{Y}_{[0,a]} \cdot \text{logit}\left(C_{[0,a]}\right) + \hat{Y}_{[a,b]} \cdot \text{logit}\left(C_{[a,b]}\right)\right|\right)$$

where the combined confidence again satisfies Assumption 4 (calibration).

See supplement for the proof. In our numerical example, the evidence from the first interval is $E_{[0,a]} = Y \cdot \hat{Y}_{[0,a]} \cdot \text{logit}\left(C_{[0,a]}\right) = +1.1$. The evidence from both intervals combined is $E_{[0,b]} = Y \cdot \hat{Y}_{[0,b]} \cdot \text{logit}\left(C_{[0,b]}\right) = +2.2$. Thus, it is straightforward to see that the evidence from the second interval is $E_{[a,b]} = E_{[0,b]} - E_{[0,a]} = 2.2 - 1.1 = +1.1$.

$$\begin{aligned} &E_{[0,a]} = +1 \cdot +1 \cdot \text{logit}(.75) = +1.1 \\ &E_{[0,b]} = +1 \cdot +1 \cdot \text{logit}(.90) = +2.2 \\ \rightarrow\ &E_{[a,b]} = +1 \cdot +1 \cdot \text{logit}(.75) = +1.1 \end{aligned}$$

We thereby infer the amount of evidence that would have had to be accumulated if information integration was optimal. There are hints for this optimal information integration in humans (Bogacz, 2007; Burr et al., 2009; Ernst & Banks, 2002). But this is clearly not always the case; integration of different stimuli or modalities is often suboptimal (Adler & Ma, 2018; Rahnev & Denison, 2018; West et al., 2023), potentially because internal uncertainty estimation is not perfect but relies on heuristic computations (Aitchison et al., 2015; Bertana et al., 2021; West et al., 2025). Crucially, our approach to measuring accumulated evidence cannot differentiate between a situation in which little evidence is integrated optimally or more evidence is integrated suboptimally. Thus, the inferred $E_{[a,b]}$ should be interpreted as the evidence that is additionally accumulated, which is not necessarily the full amount of evidence processed because some of it may have been lost in imperfect integration. (Note that other measures, e.g., accuracy and d', also do not solve this ambiguity.) With this caveat, it is nevertheless instructive to make use of additivity in order to directly measure accumulated evidence.

This numerical example demonstrates additivity of evidence in a single trial. But because measuring $E_{[0,a]}$ and $E_{[0,b]}$ is practically impossible within one trial without interrupting the perceptual processes, we next turn to computing averages across multiple trials, some with a short and others with a long presentation duration. There, additivity is preserved, and measuring evidence in the way suggested here fulfills the intuitive desideratum that—on average—multiple independent observations of the same distribution yield linearly compounding evidence.

1.2. Additive evidence averaged across trials

We illustrate the additivity of averaged accumulated evidence in a simple coin toss problem: There are two biased coins, one is biased towards heads (75% heads, 25% tails) and the other towards tails (25% heads, 75% tails). One coin is randomly chosen with equal probabilities. This coin is then flipped a number of times, n. From the outcomes of these flips, an ideal observer makes guesses about which coin was chosen. The performance of the ideal observer is shown in Fig. 1 for different ways of measuring performance.

Intuitively, we want a performance measure in this scenario to increase on average by some constant amount for each independent coin flip. Only the evidence measured as defined above fulfills this desideratum as shown in Fig. 1e. Other measures do not show a linear increase. For example, the accuracy (Fig. 1a) necessarily cannot be linear as it is bounded by 1. Computing the logarithmic odds of the

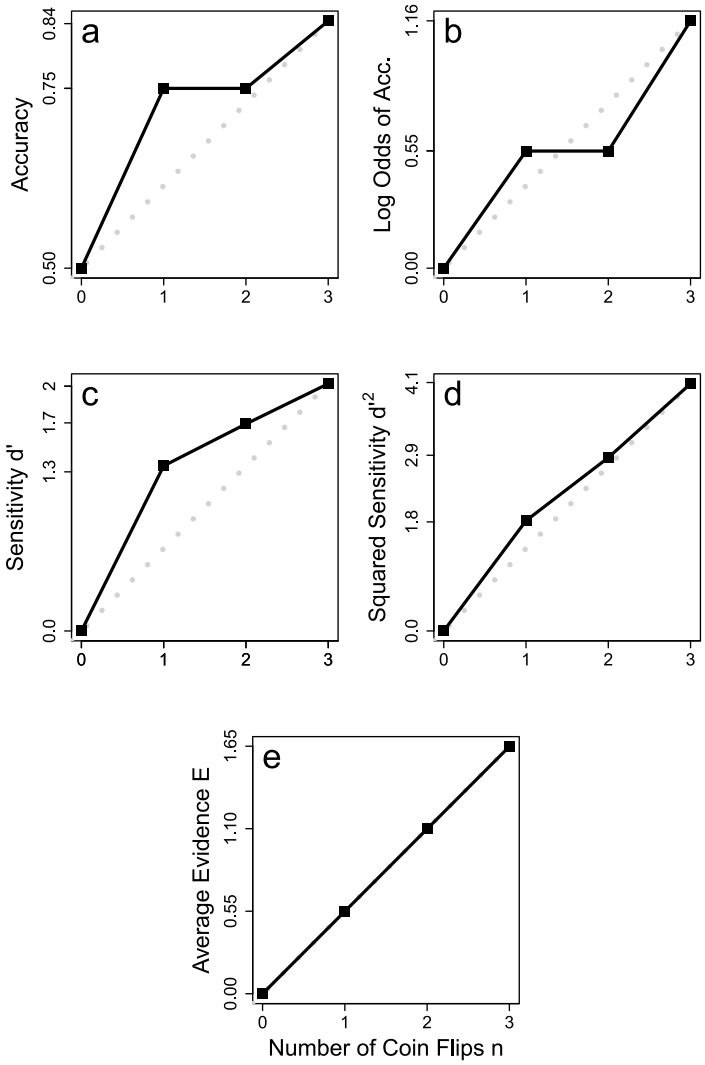


Fig. 1. Coin flip example.

Note. A randomly chosen coin being either heads-biased (75% heads, 25% tails) or tails-biased (25% heads, 75% tails) is flipped a number of times (x-axis). The number of observed coin flips determines the performance of an ideal observer guessing which coin was chosen. Measuring this via accuracy (subplot a), logarithmic odds of the accuracy (b), sensitivity d' (c), or squared sensitivity d'^2 (d) does not provide a linear increase as desired. Only average evidence (e), defined as mean logarithmic odds of posterior probabilities, produces the desired linear increase represented by the diagonal dotted line.

accuracy (Fig. 1b) does not fix this because individual confidence ratings are not considered. Signal detection theory's sensitivity d' (Fig. 1c), although unbounded, does not show a linear increase of evidence either and neither does its squared counterpart d'^2 (Fig. 1d), which only shows linear increases when the normality assumption holds—which is not the case in the coin toss example.

To understand this behavior, consider the case that the coin flips heads. The posterior probability for the chosen coin being the heads-biased coin is

$$P(\text{heads coin} \mid \text{heads flipped}) = \frac{P(\text{heads flipped} \mid \text{heads coin})P(\text{heads coin})}{P(\text{heads flipped})} = .75.$$

Thus, the ideal observer would guess the heads-biased coin, $\hat{Y} = +1$, with confidence rating $C = 75\%$. If the chosen coin is indeed the heads-biased one ($Y = +1$), the evidence is positive,

$$\begin{aligned} E &= Y \cdot \hat{Y} \cdot \text{logit}\,(C) \\ &= +1 \cdot +1 \cdot \text{logit}\,(.75) \\ &= +1.1, \end{aligned}$$

as shown in the first row of Table 1. But if the tails-biased coin was chosen ($Y = -1$) and misleadingly flipped heads, the evidence would be negative, $E = -1.1$, second row of Table 1. Analogously, flipping tails could provide positive or negative evidence depending on whether it veridically came from the tail-biased coin or not. Since the positive evidence cases are more probable, the average evidence amounts to a positive value $\bar{E} = (.375 + .375) \cdot (+1.1) + (.125 + .125) \cdot (-1.1) = 0.55$.

Thus, a single coin flip yields an average evidence of $\bar{E} = 0.55$. Repeating the same calculations for multiple coin flips yields an average evidence of $\bar{E} = 1.10$ for two flips, $\bar{E} = 1.65$ for three flips, and in general for n coin flips $\bar{E} = 0.55 \cdot n$. This linear trajectory is shown in Fig. 1e. This additivity holds in any numerical example and any scenario: For any underlying distribution and as long as we have access to the posterior probabilities (which we will approximately recover from participants' calibrated confidence ratings in our experiments), evidence will exhibit this desirable behavior.

One may object that replacing the normal noise assumption by using confidence ratings is throwing the baby out with the bath water. After all, when assuming normal noise, we know from the cue combination literature (Burr et al., 2009; Ernst & Banks, 2002) that the squared

Table 1
Coin flip example.

Coin	Outcome	Probability $P(\text{Coin}, \text{Outcome})$	Evidence
Heads-Biased	Heads	.5 · .75 = .375	+1.1
Tails-Biased	Heads	.5 · .25 = .125	−1.1
Heads-Biased	Tails	.5 · .25 = .125	−1.1
Tails-Biased	Tails	.5 · .75 = .375	+1.1
		Average evidence $\bar{E} = +0.55$	

Note. All combinations of the chosen coin and outcomes (heads or tails) together with their joint probability and the evidence these cases provide. Evidence is coded such that positive evidence indicates correct and negative evidence incorrect decisions about the coin.

sensitivity d'^2 is indeed additive: When an ideal observer combines two independent observations from the two intervals $[0, a]$ and $[a, b]$, squared sensitivities add up. With the squared sensitivity, the desirable property is therefore already achieved, $d'^2_{[0,a]} + d'^2_{[a,b]} = d'^2_{[0,b]}$. This is shown in the blue line in Fig. 2d, where we simulated data with draws from underlying normal distributions, which translates into a particular assumption on the distribution of confidence ratings. There, the d'^2-approach produces a linear trajectory reflecting the correct interpretation that the two draws yield equivalent information.

However, this approach relies on the normal noise assumption and interpretations can be biased if it is violated. For example, Lévy flight models (Voss et al., 2019) posit internal noise distributions with heavier tails all the way to the Cauchy distribution, which seem to better explain the data in some cases with speeded tasks (Rasanan et al., 2023; Voss et al., 2019; Wieschen et al., 2020) but not necessarily in others. When the underlying noise deviates from normal noise, or even when comparing conditions with different underlying noise distributions, evidence accumulation results based on d'^2 will be biased. In our simulations with underlying Cauchy distributions, the d'^2-approach produces a deviation from linearity, see red dotted line in Fig. 2d. It appears as if the evidence accumulation rate decelerates even though, again, the two draws were equally informative. In contrast, our approach solves this and produces a linear increase independent of the underlying distribution as shown in Fig. 2e.

1.3. Confidence calibration

Forgoing the normal noise assumption, we instead rely on confidence ratings to define the measure of accumulated evidence. For this, we have to ensure that Assumption 4 requiring calibrated confidence ratings roughly holds. We do so by translating participants' raw confidence ratings $\tilde{C}$ into calibrated confidence ratings, C. Calibration is done based on Zhang and Maloney (2012; Zhang et al., 2020): Perhaps not surprisingly, Zhang and Maloney demonstrated a linear relationship between raw confidences $\tilde{C}$ and accuracy (the probability of correct predictions) when both are logarithmic-odds transformed. Deviations from calibration can therefore be expressed by linear regression coefficients in the logarithmic odds space,

$$\log\left(\frac{\text{acc}}{1-\text{acc}}\right) = (1-\beta)\log\left(\frac{\alpha}{1-\alpha}\right) + \beta\log\left(\frac{\tilde{C}}{1-\tilde{C}}\right)$$

where the regression intercept α corresponds to a specific confidence to which participants' ratings are biased and the weight β (functioning as a regression slope) indicates the strength of that bias.

For estimating these coefficients, we sorted all trials by the raw confidence ratings $\tilde{C}$ into quartile bins (H_i, $i \in \{1, 2, 3, 4\}$) of equal size. For each bin, we computed the logarithmic odds of the accuracy and the average logarithmic odds of confidence ratings. We then fitted the parameters of the linear regression ($\hat{\alpha}$, $\hat{\beta}$). With that, we inverted the linear model to obtain a function that allows calibrating raw confidence ratings in each trial, $\tilde{C}$, to the calibrated confidence rating, C.

$$C = \frac{1}{1 + e^{-\left((1-\hat{\beta})\log\left(\frac{\hat{\alpha}}{1-\hat{\alpha}}\right) + \hat{\beta}\log\left(\frac{\tilde{C}}{1-\tilde{C}}\right)\right)}}$$

This ensures that confidence ratings are roughly calibrated as posited in Assumption 4.

Note that informative confidence ratings increase the measured evidence in our approach. If participants gave raw confidence ratings entirely unrelated to the quality of their decisions, this procedure would map these raw confidence ratings all to the same constant value (equal to the overall accuracy). In this worst case, calibration is still trivially achieved but, because the logarithmic odds is a convex function in the positive range, this would yield lower evidence values compared to when participants give informative confidence ratings.

We applied this calibration separately for each participant and, within each participant, separately for each session and condition. Separating calibration by condition is important because confidence ratings can contain biases between conditions (Rahnev & Denison, 2018), which would then translate into biased estimates for calibrated confidence ratings. To avoid these biases, it is essential for the validity of our evidence measure to perform calibration within conditions. We elaborate on this issue as a limitation in the General Discussion.

Differences from DDMs

Before continuing, it is important to distinguish our approach from that of DDMs. Both are grounded on the principles of SPRT. But DDMs consider evidence as a latent variable that accumulates internally until a threshold is reached and a decision is made by the participant. This way, DDMs model decision and response time distributions. In contrast, we use participants' decisions and confidence ratings and transform them into a performance measure—akin to accuracy or d'—which we call accumulated evidence. This is not a latent variable anymore but an observable measure. Instead of response time, we consider how stimulus presentation time affects evidence accumulation. For this purpose, we will plot stimulus presentation time (x-axis) against the accumulated evidence computed from participants' responses (y-axis) in subsequent plots. Note that we also do not use DDM-variants that explain confidence ratings (Glickman et al., 2022; Hellmann et al., 2024; Lee et al., 2023; Moran et al., 2015; Pleskac & Busemeyer, 2010) but only build our measure based on the given confidence responses.

1.4. Simulations based on meta-d'

To validate our method, we conducted simulations based on the model that underlies what is considered the gold standard in metacognitive research (Fleming, 2024; Katyal & Fleming, 2024; Michel, 2022): metacognitive sensitivity, meta-d', and metacognitive efficiency, M-ratio = meta-d'/d' (Maniscalco & Lau, 2012, 2014). These two measures evaluate the ability to give appropriate confidence ratings on the same metric as sensitivity d'. In the underlying model, a first decision variable is drawn based on underlying normal noise with sensitivity d'. Then, a second, metacognitive decision variable is drawn which is based on independent truncated normal noise with metacognitive sensitivity meta-d' (Rausch et al., 2023). The metacognitive sensitivity meta-d' can be higher or lower than d'. The mismatch is reflected in the measure of metacognitive efficiency, M-ratio = meta-d'/d'. As with signal detection theory, the sampled, metacognitive decision variable is often assumed to be thresholded into different confidence rating levels.

Fig. 3 shows how our approach evaluates evidence in this model. For simplicity and for consistency with the previous examples, we simulated data with a sensitivity of $d'_{[0,a]} = 1.3$ in the short and a sensitivity of $d'_{[0,b]} = 1.9$ in the long condition (corresponding to two independent, equivalent draws with $d'_{[0,a]} = d'_{[a,b]} = 1.3$ which combine into $d'_{[0,b]} = 1.9$). For simulating confidence ratings, we varied M-ratio in $\{0.5, 0.75, 1\}$. This is the range of values that is most often encountered in empirical data reflecting additional noise in the confidence ratings (see e.g., Sadnicka et al., 2024). For each simulated participant, we generated 80 trials in the short and 80 trials in the long condition. We show standard errors that would be expected for 10 participants (assuming no variability between participants). This equals the lowest

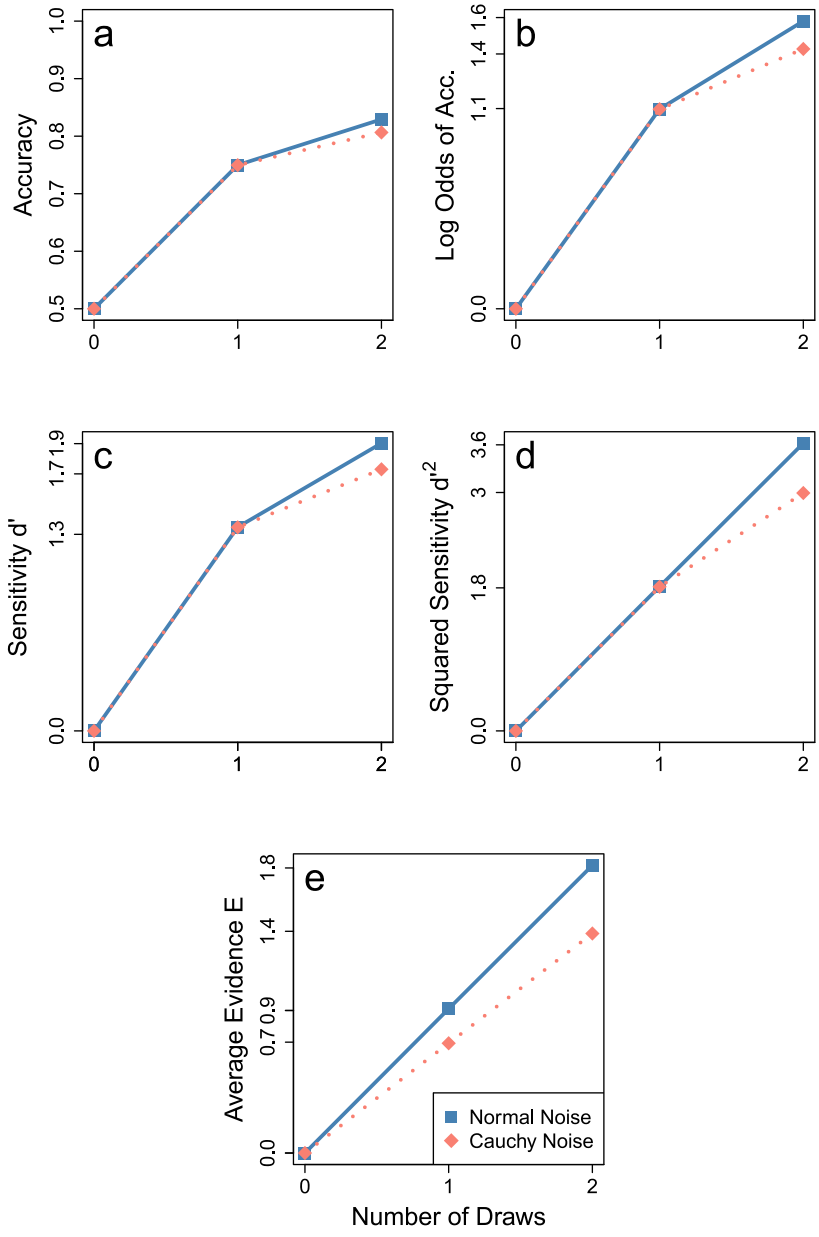


Fig. 2. Evidence accumulation with normal and cauchy noise.

Note. Similar to Fig. 1, an ideal observer's performance is evaluated for up to two draws (x-axis) in either a signal detection theory setting with normally distributed decision variables (blue squares) or when replacing the normal distributions with Cauchy distributions (red diamonds). The separation of distributions in the two cases is equated such that a single draw produces an accuracy of 75%. However, because the ideal combination of Cauchy draws is less informative, all performance measures are lower when combining two such draws than in the normal case. Again, measuring accuracy (a), logarithmic odds of accuracy (b), sensitivity (c), and squared sensitivity (d) do not yield a linear increase. Only the average evidence (e) shows this desirable property.

number of participants and trials in any of the calibration conditions we investigated in our experiments.

We additionally varied how the resulting underlying metacognitive decision variables translate into reported confidence ratings. In Fig. 3a, we simulated a situation in which the (noisy) metacognitive decision variables directly translate into the confidence ratings, which are then calibrated as discussed above. M-ratio = 1 (uppermost line) is the special case corresponding to what we have shown in Fig. 2e already. As before, accumulated evidence starts to rise from 0 (before the stimulus is shown) and rises linearly. This climb is slower with additional noise on confidence ratings (M-ratio < 1; lower two lines in Fig. 3a). This is expected because confidence ratings are less informative. Although there is overall less evidence, trajectories remain near-linear for these M-ratio values, correctly reflecting the constant accumulation of evidence.

In contrast, the linearity of the accumulation process is misjudged by our approach when confidence responses are discretized through metacognitive thresholds even when calibrated. This is because confidence rating distributions move from low to high throughout the accumulation process, and high evidence values are produced when these thresholds are met. Thus, if thresholds are set to differentiate between confidence ratings at relatively low levels, more evidence appears to be accumulated in the initial rather than later stages, see Fig. 3b. But the reverse is also possible. If metacognitive thresholds

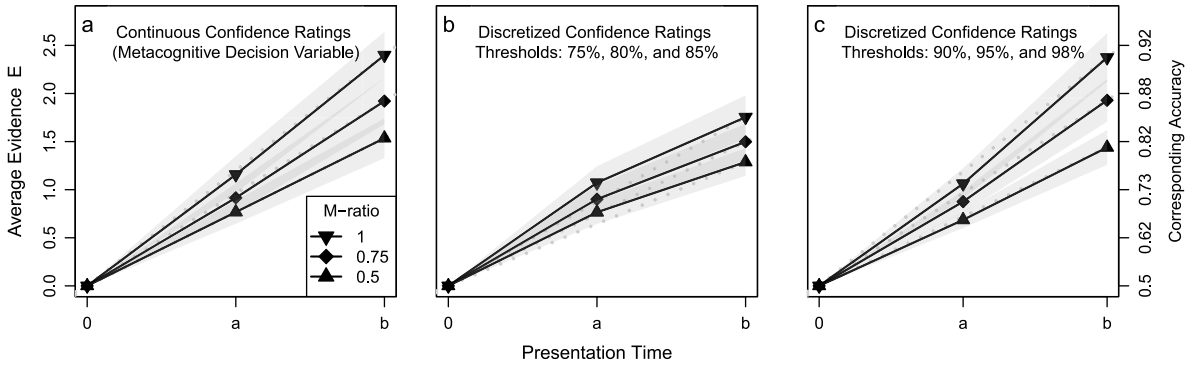


Fig. 3. Simulations based on the model underlying meta-d' and M-ratio.
Note. We simulated data based on the meta-d' model with constant underlying evidence accumulation with sensitivities $d'_{[0,a]} = 1.3$ and $d'_{[a,b]} = 1.3$ and different M-ratios (varied between $\{0.5, 0.75, 1\}$). (a) If participants' confidence ratings are proportional to the metacognitive decision variable in that model (without metacognitive thresholds), evidence accumulation trajectories are approximately linear (dotted lines). (b) But deviations occur when confidence ratings are discretized. If metacognitive thresholds differentiate well between low confidence levels, evidence accumulation appears to be better in the first than in the second presentation time window—an apparent deceleration of evidence accumulation. (c) If they differentiate between high confidence levels, the pattern is reversed—an apparent acceleration. Thus, discretizing confidence ratings can lead to inappropriate conclusions about the underlying accumulation trajectories. Shaded areas correspond to expected standard errors for $n = 10$ participants with 80 trials in the short ($[0, a]$) and 80 trials in the long ($[0, b]$) condition.

Table 2
SEM depending on d' and trial numbers.

$d'_{[0,a]}$	Trial number	$SEM_{E_{[0,a]}}$
0.5	40	0.10
0.5	80	0.04
0.5	160	0.03
0.5	320	0.02
1.0	40	0.22
1.0	80	0.12
1.0	160	0.06
1.0	320	0.04
2.0	40	0.31
2.0	80	0.25
2.0	160	0.19
2.0	320	0.14

Note. We estimated standard error of means (SEM) of $E_{[0,a]}$ for $n = 10$ participants, different sensitivities $d'_{[0,a]} \in \{0.5, 1.0, 2.0\}$, M-ratio = 1 (other values showed somewhat similar results), and different trial numbers per participant (varied within $\{40, 80, 160, 320\}$).

are set to differentiate between confidence ratings at relatively high levels, more evidence appears to be accumulated later, see Fig. 3c. Therefore, we discourage using our method with discretized confidence rating categories where participants are forced to set metacognitive thresholds.

For empirical applications, obtaining reliable estimates is crucial. To validate the calibration process in simulations and get a grasp on how many trials are required, we show simulations based on the meta-d' model from Fig. 3a in Fig. 4. To translate these results into more fine-grained recommendations on the number of required trials, we further computed the standard errors of the accumulated evidence in Table 2. For example, with an expected $d'_{[0,a]} = 1.0$ which would translate to $E_{[0,a]} = 0.5$ and 80 trials, the resulting standard error is approximately $SEM = 0.12$ (sixth row in Table 2). Note that larger evidence values entail larger standard errors because, when accuracies are close to 100%, calibration becomes less stable.

Thus, when assuming the standard model based on (truncated) normal distributions to produce confidence ratings, our approach can, with at least 10 participants and 80 trials per calibration condition, identify linear evidence accumulation trajectories if the underlying information process is indeed linear. With our notion of evidence based on calibrated confidence ratings and its main advantage of additivity in mind, we next turn to empirical applications of this approach.

2. Experiment 1: Effect of causality on evidence accumulation

The first experiment primarily serves as a validation for our method of measuring evidence accumulation. We presented moving dots for a duration of either $a = 3.75$ s (short condition) or $b = 7.5$ s (long condition). By presenting only five dots (as opposed to many more in the typical random-dot kinematograms, see Experiment 2), we increased sensory evidence although categorical evidence remains low, which is expected to provide a relatively constant evidence accumulation rate throughout stimulus presentation (Lange et al., 2021).

In addition, we adapted the stimuli from Meding et al. (2019) to present dot motion in two conditions: causal and anti-causal. This allowed us to test whether the visual system accumulates more evidence from stimuli being presented in a causal rather than an anti-causal motion (dot motion that is played backward), see below for details.

2.1. Method

The experiment was conducted in accordance with the 1964 Declaration of Helsinki, with the ethical guidelines of the German Psychological Society (DGPs), and with the Professional Association of German Psychologists (BDP, 2005, C.III). The experiment was approved by the local ethics committee (reference number 2022_0127_246).

The experiment code was written in MATLAB using the Psychophysics Toolbox (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997). The experimental MATLAB code, data, and analyses in R are freely available at osf.io/krw53.

Participants

Eleven participants (aged 18–21, ten right-handed and one left-handed, seven female and four male) took part in the experiment. Participants were cognitive science students from the University of Tübingen, Germany, naive to the purpose of the experiment, had normal or corrected-to-normal vision, and received either course credit or monetary compensation (10 Euros per hour).

The number of participants (and trials) was not preregistered. In a prior power analysis, we reasoned that—if all but one participant showed the effect in the anticipated direction (i.e., more evidence

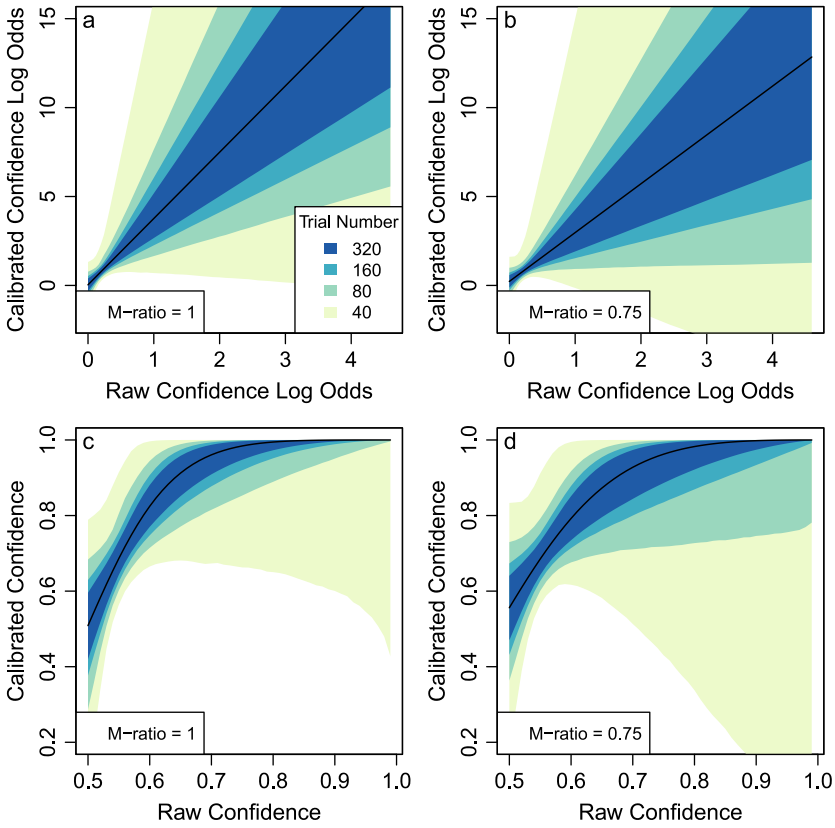


Fig. 4. Reliability of logarithmic odds calibration.
Note. The reliability of the calibration process based on logarithmic odds depends on the number of trials. For simplicity, we base simulations again on the meta-d' model without metacognitive thresholds (as in Fig. 3a). Here, we show calibration reliability for a single simulated participant and only for the first presentation interval ($d'_{[0,a]} = 1.3$) for M-ratio = 1 (a and c) and M-ratio = 0.75 (b and d). The black lines show the perfect calibration that would ideally be applied to participants' responses. Shaded areas show the 95% confidence intervals for regression lines for different numbers of trials. Although the confidence intervals around regression lines fan out considerably on the logarithmic odds scale (a and b), when transformed back to the confidence scale via logit^{-1} (c and d), variability is sufficiently controlled.

accumulation in causal versus anticausal condition)—a binomial test would yield sufficient evidence with nine participants. Rounding up, we set out to sample 10 participants but, due to overbooking, we invited 11 participants. Importantly, each participant took part in three sessions on separate days so that we could—consistent with psychophysical tradition (Normand, 2016; Rouder & Haaf, 2018; Smith & Little, 2018)—measure each individual with high precision. This was also necessary to ensure a reliable calibration procedure.

Procedure

At the start of the experiment, participants were informed about the experimental procedure and signed consent forms. They were then seated in front of a monitor (VIEWPixx /3D Lite, VPixx Technologies Inc., Montreal, Canada; 1920 × 1080 pixels, 521 × 293 mm size, 120 Hz refresh rate) at a distance of approximately 60 cm in a sound-proof cabin. Participants conducted two practice blocks followed by eight experimental blocks. Practice blocks contained four trials each and experimental blocks contained 32 trials each for a total of 256 experimental trials per session. Participants were encouraged to take self-paced breaks between blocks.

This procedure was repeated for three sessions resulting in a total of $256 \cdot 3 = 768$ experimental trials per participant. Each session took place on a different day with one to 13 days in between sessions. There were no dropouts. Sessions lasted between 1.25 to 2.25 h.

Stimuli

In each trial, participants were presented with five dots stochastically moving left or right. We primarily manipulated the presentation duration of the dot motion: We displayed them either for a short duration (15 motion steps, each frame displayed for 250 ms) or a long duration (30 motion steps), see Fig. 5. Dots were white (103 cd/m^2) with a diameter of 1.43 degree of visual angle (deg) and blurred with a Gaussian standard deviation of 0.24 deg (matching those of Meding et al., 2019). These dots were presented on a black (0.08 cd/m^2) background.

The motion of these dot stimuli was inspired by Meding et al. (2019). The main difference to their experiment was that we used five independently moving dots instead of only one. We determined the position of the dots in each frame t by an auto-regressive process, $x_t = 0.4 \cdot x_{t-1} + 0.2 \cdot x_{t-2} + 0.1 \cdot x_{t-3} + 0.05 \cdot x_{t_4} + \epsilon_t + \delta$ (starting at $x_t = 0$ for $t \leq 0$). Meding et al. (2019) used the same auto-regressive process but without drift δ because they were interested in whether participants could discriminate the stimuli to be causal (sequence of dot positions at $x_0, x_1, \ldots, x_{n-1}, x_n$) or anti-causal ($x_n, x_{n-1}, \ldots, x_1, x_0$). These two presentations are statistically indiscriminable if ϵ_t is normally distributed. Therefore, Meding et al. (2019) chose a different noise distribution: $\epsilon_t = \text{sign}(Y_t) \cdot |Y_t|^r$ with Y_t being normally distributed. In this setting, r determines the strength of the deviation from normal noise: the further r from 1, the higher the discriminability between both presentation orders. In our experiment, we used $Y \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 65.2$ deg and $r = 0.5$ leading to a bimodal noise distribution with a standard deviation of ϵ_t of around 1.16 deg. At $r = 0.5$, Meding et al. (2019) found participants to be able to discriminate between causal and anti-causal motion with an accuracy of around 80%. Thus, we considered this a good intermediate condition to test our hypothesis:

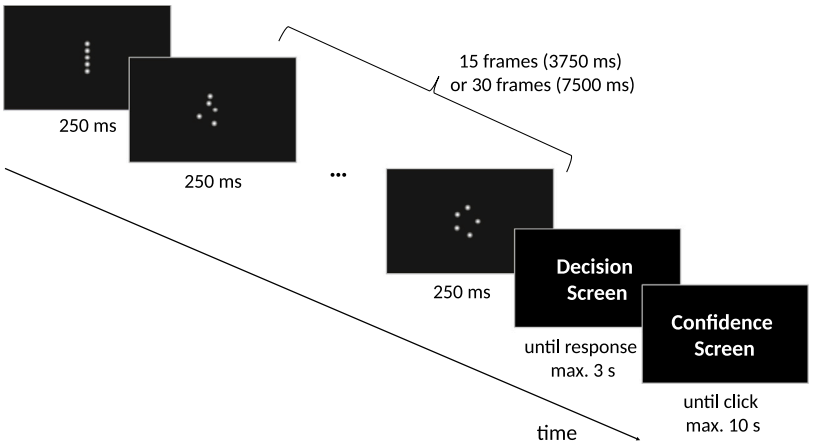


Fig. 5. Trial structure in Experiment 1.
Note. All trials begin with one frame of five dot stimuli vertically aligned followed by a sequence of either 15 (short condition) or 30 frames (long condition) in which the dots moved stochastically to the left or to the right. After that, participants responded to the motion direction of the stimuli (left or right) and provided confidence ratings.

In our experiment, participants did not actively discriminate causal versus anti-causal motion but had to discriminate the motion direction. For that, we added the aforementioned drift δ to induce leftward ($\delta = -0.11$ deg) or rightward motion ($\delta = +0.11$ deg, which we determined to achieve intermediate stochastic differentiability between the motion).

To control the amount of evidence we displayed to participants about the motion, we used a frozen-noise approach: Instead of randomly generating new stimuli for each trial (with new ε_t), we simulated one long, rightward, causal motion ($x_0, x_1, x_2, \ldots, x_{15}, x_{16}, \ldots x_{30}$) by concatenating two short motion paths to ensure that both intervals are equally discriminable despite the auto-regressive process being weakly stationary. From this, we generated the other stimulus conditions preserving the informativeness of the steps. In total, there were eight conditions corresponding to the cross-combinations of {short, long} × {causal, anti-causal} × {left, right}. For the short condition, we simply truncated the sequence to ($x_0, x_1, x_2, \ldots x_{15}$). To flip the causal direction, we inverted the order of the motion differentials in both intervals separately: Starting with aligned positions, the motion differentials were applied in the inverse order. Motion direction was flipped from rightward to leftward motion by mirroring dot motion in the middle of the screen. This way, we generated stimuli with matched discriminability in the eight conditions. We generated 32 such matched stimulus groups leading to a total of $32 \cdot 8 = 256$ trials per session, which were then repeated in each session in a randomized order.

Participants then judged whether the motion was to the left or to the right. If the response took longer than 3 s, the trial was discarded and repeated again at a random point within the remainder of the current block. If the response was in time, participants proceeded to give their confidence rating for which they similarly had a maximum response time of 10 s.

Confidence ratings

Participants gave confidence ratings on a visual analogue scale with logarithmic odds scaling. This was inspired by Phillips and Edwards (1966) who found better calibration in participants' confidence ratings using this type of scale. We modified this scale slightly to go from 50% (guessing) to 99% (almost certain) with accelerating changes in confidence (smaller distance between 50% to 55% than between 90% and 95%) following a logarithmic odds scaling. The numerical value of the currently selected confidence rating was dynamically presented, which Matejka et al. (2016) found to produce the best compromise of unbiasedness and precision among visual analogue scale variants. Participants were instructed to match their confidence ratings to the probability of their responses being correct (i.e., give 70% confidence rating if 70% of trials with that confidence rating are correct).

Participants were monetarily incentivized to give appropriate raw confidence ratings in this task by receiving up to 2 Euros per session. Such incentives need to be treated with caution because they can influence participants' metacognitive thresholds to maximize reward with discrete ratings (Maniscalco et al., 2025). Here, with continuous confidence ratings, we employed a method that prevents basic exploitation strategies to maximize reward without evaluating uncertainties. The incentive after each session was calculated based on the expected calibration error, ECE (Guo et al., 2017). For this, we binned raw confidence ratings into 10 bins, $[0\%, 10\%)$, $[20\%, 30\%)$, $\ldots$, $[90\%, 100\%)$. (Bins with confidence ratings below 50% are relevant when we later evaluate ECE after calibration: Even though participants cannot give raw confidence ratings below 50%, calibration can produce such values if participants' low confidence ratings are more often incorrect than correct.)

We computed mean differences between the accuracy and the averaged, raw confidence rating in these bins,

$$ECE = \sum_{m=1}^{M=10} \frac{|B_m|}{n} |acc(B_m) - c(B_m)|$$

where, by convention, we set the term in the sum to 0 if no responses were given in the corresponding bin ($|B_m| = 0$). This ECE takes values between 0 and 1. The lower the ECE, the better the confidence is calibrated. Hence, the incentive computation is based on its complement, $1 - ECE$. Besides calibration, we also had to take accuracy into account because, otherwise, participants could have trivially maximized their incentive by always deciding randomly and giving a 50% confidence rating. Thus, we computed the session-wise incentive as Incentive $= (0.7 \cdot (1 - ECE) + 0.3 \cdot \text{accuracy}) \cdot 2$ Euros. To maximize their payout, participants had to give accurate responses but also estimate their uncertainty. But note that, if a participant obtained an estimate of their overall accuracy in this experiment, they could always report that value rather than varying their confidence rating based on their uncertainty from trial to trial. This strategy would not be penalized by our incentive—but we considered this a negligible possibility here.

2.2. Results

Participants were able to discriminate the motion direction with an accuracy of 67% (95% CI $[64, 71]$) in the short condition and a higher

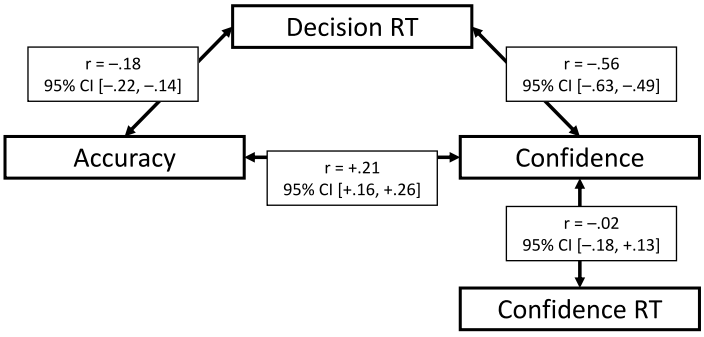


Fig. 6. Correlation of main variables in Experiment 1.
Note. For each participant, a correlation was computed between the accuracy (percentage of correct responses over all conditions), decision response times (RT), confidence ratings, and response times in giving these confidence ratings. These individual correlations were Fisher z-transformed (Fisher, 1915) to compute an average and confidence intervals, which were then transformed back to the correlation values shown here.

accuracy of 76% (95% CI [72, 79]) in the long condition. The difference in accuracy was 8%-points, 95% CI [6, 10]. Participants also showed a higher accuracy in the causal (74%, 95% CI [69, 78]) than in the anti-causal condition (69%, 95% CI [66, 72]), with a difference of 5%-points, 95% CI [1, 8].

Within participants, accuracy correlated on average negatively with the response time to decide for the stimulus direction (decision RT), see Fig. 6. Accuracy also correlated positively with confidence ratings, as expected. Moreover, confidence ratings showed a clear negative correlation with decision RT but not with the time to give the confidence rating itself (confidence RT). These correlations are in line with previous, established results (Lee et al., 2023; Moran et al., 2015; Pleskac & Busemeyer, 2010) and, therefore, validate our experimental manipulation. However, the negative relationship between confidence and decision RT was substantially stronger ($r = -.56$) than what is found on average across different paradigms ($r = -.24$, Rahnev et al., 2020).

Calibration

Without calibration, participants showed marked differences in the relation between their confidence ratings and the corresponding accuracies, see Fig. 7. Most participants were overconfident (values below the diagonal). After calibration, these deviations were mostly eliminated supporting Zhang and Maloney (2012)'s notion of a linear relationship of confidence ratings and accuracy in the logarithmic odds space.

Evidence accumulation

Evidence accumulation was approximately linear (although not perfectly so) with surprisingly stable trajectories despite large individual differences. Fig. 8 shows the mean evidence for each participant for short and long conditions: Trajectories are roughly on a line so that mean evidence in the short presentation duration ($\bar{E}_{[0,a]} = 0.48$, 95% CI [0.34, 0.62]) almost doubled in the (twice-as-)long presentation duration ($\bar{E}_{[0,b]} = 0.86$, 95% CI [0.61, 1.11]). However, evidence did not exactly double. There were some deviations from linearity, which became apparent when computing the average evidence accumulated exclusively in the second interval of stimulus presentation, $\bar{E}_{[a,b]} = \bar{E}_{[0,b]} - \bar{E}_{[0,a]}$. Here, the average accumulated evidence was $\bar{E}_{[a,b]} = 0.38$, 95% CI [0.26, 0.51]. With that, there was less evidence accumulation in the second half of stimulus presentation than in the first with a difference of $\bar{E}_{[0,a]} - \bar{E}_{[a,b]} = 0.10$, 95% CI [0.03, 0.17], see Fig. 8a.

Note that testing for a difference across participants can be misleading: Participants with an overall high rather than low evidence accumulation receive more weight in this analysis. Therefore, we also conducted a ratio test computing $\bar{E}_{[a,b]}/\bar{E}_{[0,a]}$ for all participants and testing against the ratio 1 which would reflect a constant evidence accumulation rate. Conversely, this analysis comes with its own problem: Now participants with an overall low evidence are weighted more, and the same absolute amount of noise can create large fluctuations, especially when the denominator $\bar{E}_{[0,a]}$ is close to 0 (Franz, 2007; Von Luxburg & Franz, 2009). Reassuringly, we obtained a very similar result: A decrease in evidence accumulated in the second compared to the first interval with a mean ratio of $M = 0.8$, 95% CI [0.6, 1.0]. Given that both analyses yielded the same result and that this pattern was stable across participants, we conclude that evidence accumulation in this experiment was slightly higher in the first interval of presentation duration as compared to the second.

Regarding causality, we had already reported a higher accuracy in the causal than anti-causal condition. In terms of evidence, we found the same result, see Fig. 8b: Accumulated evidence was higher in the causal than in the anti-causal condition. This was the case after a short presentation duration ($\bar{E}^{\text{causal}}_{[0,a]} = 0.55$, 95% CI [0.38, 0.73] versus $\bar{E}^{\text{anti-causal}}_{[0,a]} = 0.41$, 95% CI [0.31, 0.51]) for a difference of $\bar{E}^{\text{causal}}_{[0,a]} - \bar{E}^{\text{anti-causal}}_{[0,a]} = 0.14$, 95% CI [0.06, 0.22] as well as after a long presentation duration ($\bar{E}^{\text{causal}}_{[0,b]} = 0.93$, 95% CI [0.65, 1.22] versus $\bar{E}^{\text{anti-causal}}_{[0,b]} = 0.79$, 95% CI [0.56, 1.02] for a difference of, again, $\bar{E}^{\text{causal}}_{[0,b]} - \bar{E}^{\text{anti-causal}}_{[0,b]} = 0.14$, 95% CI [0.04, 0.25]). The advantage of causality on evidence accumulation is therefore localized in the first interval of stimulus presentation because it did not substantially increase throughout the second presentation interval, $\bar{E}^{\text{causal}}_{[a,b]} - \bar{E}^{\text{anti-causal}}_{[a,b]} = 0.00$, 95% CI [−0.08, 0.09].

To demonstrate the relevance of measuring evidence as suggested here, we plot different performance measures as in the coin example from Fig. 1 but now for the experimental data in Fig. 9. Again, accuracy, logarithmic odds of accuracy, and sensitivity d' (Fig. 9a–c) are in principle less suitable for numerical evaluation. The squared sensitivity d'^2 (Fig. 9d) would allow such evaluations under the normal noise assumption. Curiously, this d'^2-approach points to the opposite interpretation than with evidence: Squared sensitivity (averaged across participants) is numerically smaller in the first interval ($d'^2_{[0,a]} = 1.0$, 95% CI [0.7, 1.3]) than in the second ($d'^2_{[a,b]} = 1.2$, 95% CI [0.8, 1.6]). If anything it seems to indicate an acceleration, $d'^2_{[a,b]} - d'^2_{[0,a]} = 0.26$, 95% CI [−0.10, 0.62], but not conclusively so. A ratio test yielded a similar result, $d'^2_{[a,b]}/d'^2_{[0,a]} = 1.4$, 95% CI [0.9, 2.0]. Thus, this experiment represents an example of how interpretations about evidence accumulation trajectories diverge when relying on the normal noise assumption versus using the here presented notion of evidence accumulation based on empirically observable, calibrated confidence ratings.

2.3. Discussion

We have applied the definition of evidence as the logarithmic odds of (calibrated) confidence ratings in a first experiment. Based on this definition, evidence accumulation slightly decelerated. In contrast,

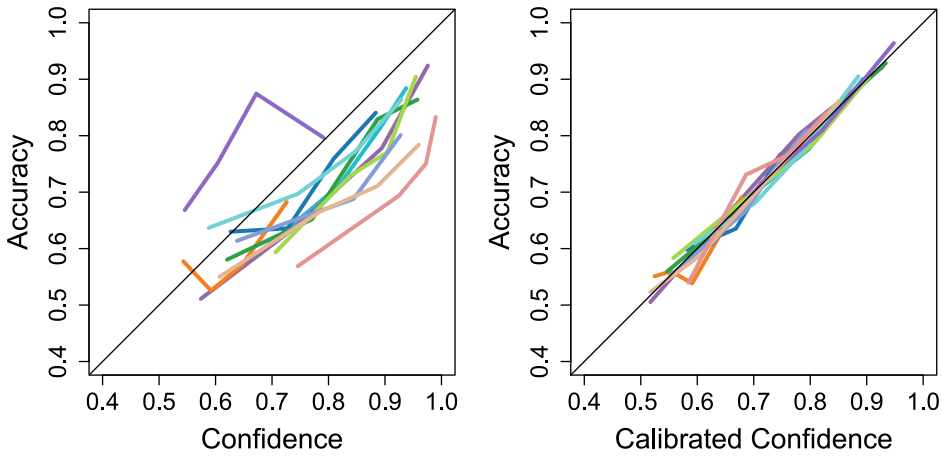


Fig. 7. Confidence calibration in Experiment 1.
Note. Participants' match between confidence ratings (x-axis) and accuracy (y-axis) before (left) and after calibration (right). For this visualization and for the calibration, each participant's confidence ratings were binned into four equally sized bins, for which the average confidence rating and accuracy were calculated. For calibration, these values were transferred into the corresponding logarithmic odds and a linear model was fit. Using this linear model, we then calibrated the logarithmic odds of the trial-wise confidence ratings and transferred them back to the trial-wise calibrated confidence ratings. Diagonal lines represent perfect calibration.

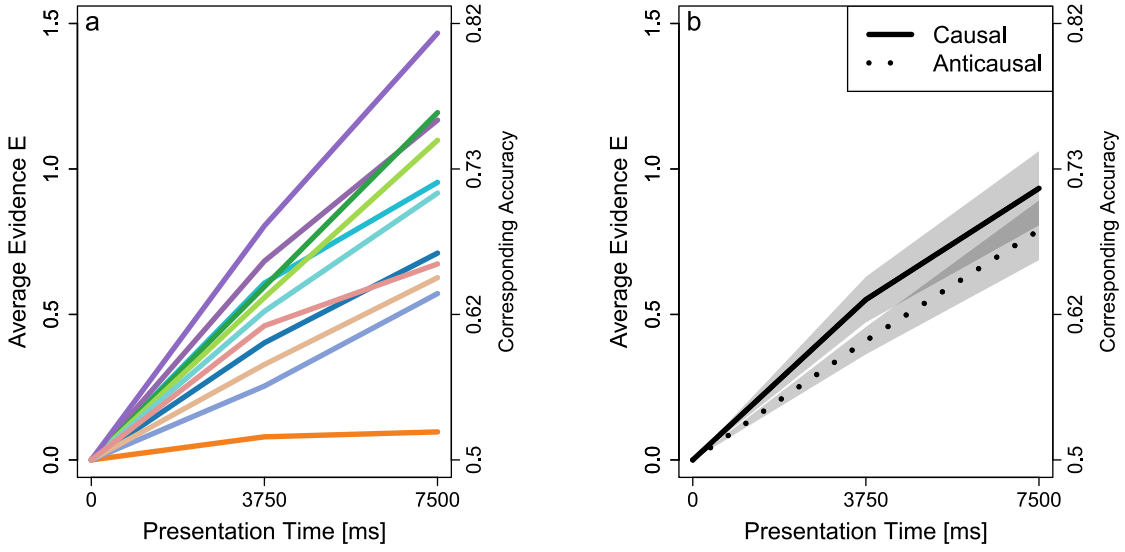


Fig. 8. Evidence accumulation in Experiment 1.
Note. Evidence accumulation in individual participants and contrasting causality conditions. (a) Each line depicts one participant's average evidence accumulation (y-axis) trajectory for short and long presentation durations (x-axis) averaged across all other conditions. The right *y*-axis shows the accuracy corresponding to the average evidence; but note that these do not necessarily reflect participants' actual accuracies in the experiment. (b) Averaging across participants but splitting by causality condition, we found higher average evidence accumulated in the causal in contrast to the anti-causal condition. Shaded bands around trajectories in (b) represent standard errors interpolated between the time steps.

assuming underlying normal noise and using squared sensitivity d'^2 as a measure of performance would have indicated the opposite effect.

This experiment used moving dot stimuli adapted from Meding et al. (2019), which allowed presenting dot motion either causally or anti-causally. Here, all performance measures agreed: More information is gathered from a causal rather than an anti-causal display. Because this is the first demonstration of this effect we are aware of and there was no preregistration of this effect, we want to interpret this observation with caution. Conceptual replications are necessary to confirm this effect. This is important because auto-regressive motion with added drift may induce statistical peculiarities that create the effect without bearing on inherent causality. More realistic causal stimuli are required to ensure the validity of this finding.

However, if the effects we found were confirmed, these results would allow for excitingly specific hypotheses: The advantage in evidence accumulation from causality was already fully present after the short presentation duration ($\bar{E}^{\text{causal}}_{[0,a]} - \bar{E}^{\text{anti-causal}}_{[0,a]} = 0.14$). In the long condition, this advantage did not further increase ($\bar{E}^{\text{causal}}_{[0,b]} - \bar{E}^{\text{anti-causal}}_{[0,b]} = 0.14$ as well) but rather plateaued: There was no substantial evidence gain from causality throughout the second interval, $\bar{E}^{\text{causal}}_{[a,b]} - \bar{E}^{\text{anti-causal}}_{[a,b]} = 0.00$. Together, this may point to motion causality being beneficial for the early stages of perceptual evidence accumulation but not later stages. In general, the ability to numerically evaluate differences with the evidence measure proposed here, allows researchers to generate more specific hypotheses.

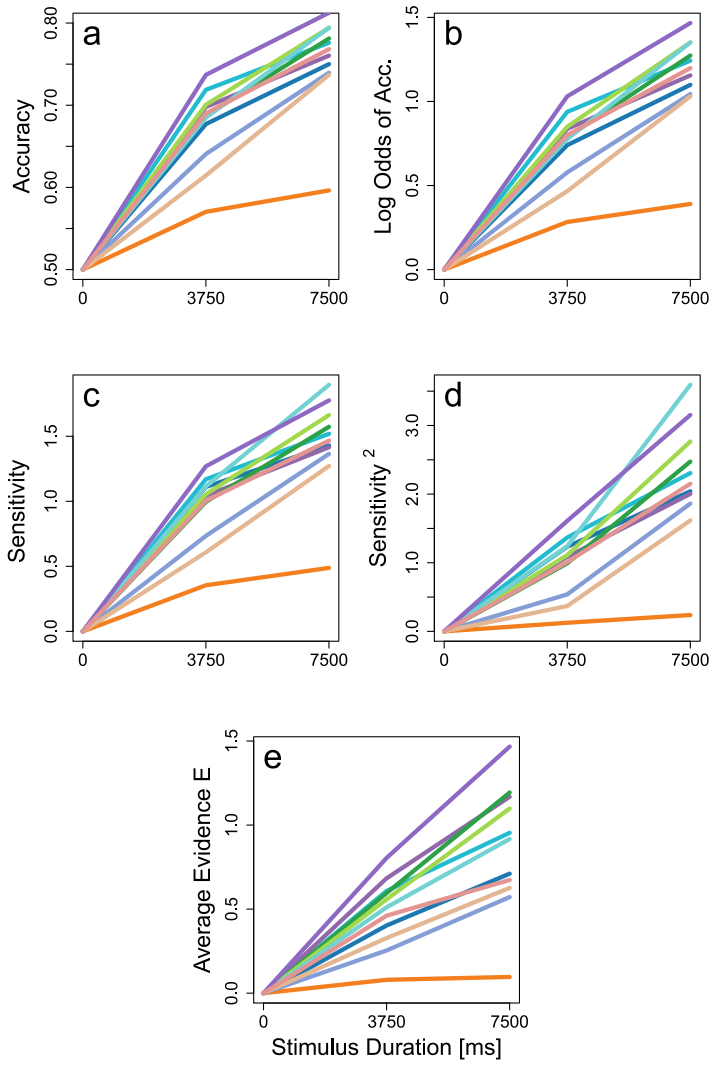


Fig. 9. Performance measures in comparison for Experiment 1.
Note. Similar to Figs. 1 and 2, the different performance measures are shown but now for the data of Experiment 1. Accuracy (subplot a), logarithmic odds of the accuracy (b), and sensitivity d' (c) show the expected decelerating trajectories. This is not surprising due to their construction as, for example, accuracy must plateau at 100%. Thus, these measures cannot be taken directly to compare numerical increments directly. In contrast, squared sensitivity d'^2 (d) allows this interpretation but requires the assumption of underlying normal noise. Average evidence (e) instead relies on empirical confidence ratings (after calibration). Curiously, squared sensitivity (d) tentatively indicates more information gathered from the second compared to the first interval of stimulus presentation. In contrast, average evidence (e) indicates the opposite.

3. Experiment 2: Dependencies during evidence accumulation

In the second experiment, we measure dependencies in the evidence accumulation. For this, we returned to the traditional random-dot kinematogram stimuli (RDK; Newsome & Pare, 1988), which are frequently used in psychophysical experiments (Desender et al., 2021; Fleming et al., 2018; Gallagher et al., 2019; Rollwage et al., 2020; Zylberberg et al., 2012, 2016): Multiple dots moved randomly with a small subset of dots moving coherently to the left or to the right. Participants had to discriminate the direction of the average motion throughout the stimulus presentation. Again, our primary manipulation was in the stimulus duration, which was either short, a = 350 ms, or long, b = 700 ms.

To investigate the dependence of evidence accumulation in the second interval of stimulus presentation (the second 350 ms interval within the 700 ms presentation duration condition), we independently manipulated the coherence (i.e., motion strength) in the two intervals to study how evidence accumulation would change, $\bar{E}_{[a,b]} \sim \bar{E}_{[0,a]}$. We also manipulated congruency of the stimulus shown in the second interval (i.e., whether the motion was the same or the opposite in the first versus second interval). This approach is conceptually similar to that of Rollwage et al. (2020) and Fleming et al. (2018), who also presented two intervals but in a temporally separated manner: Participants saw one stimulus interval, gave responses, and then saw the second interval after which they gave a second round of responses. Here, we add to their line of investigation on how evidence accumulation in the second interval is affected by the first, where our focus was on contiguous instead of temporally separated displays so as to not interrupt the perceptual process.

3.1. Method

The experiment was conducted in accordance with the 1964 Declaration of Helsinki and with the ethical guidelines of the German Psychological Society (DGPs) and the Professional Association of German Psychologists (BDP) (2005, C.III). Furthermore, the experiment was

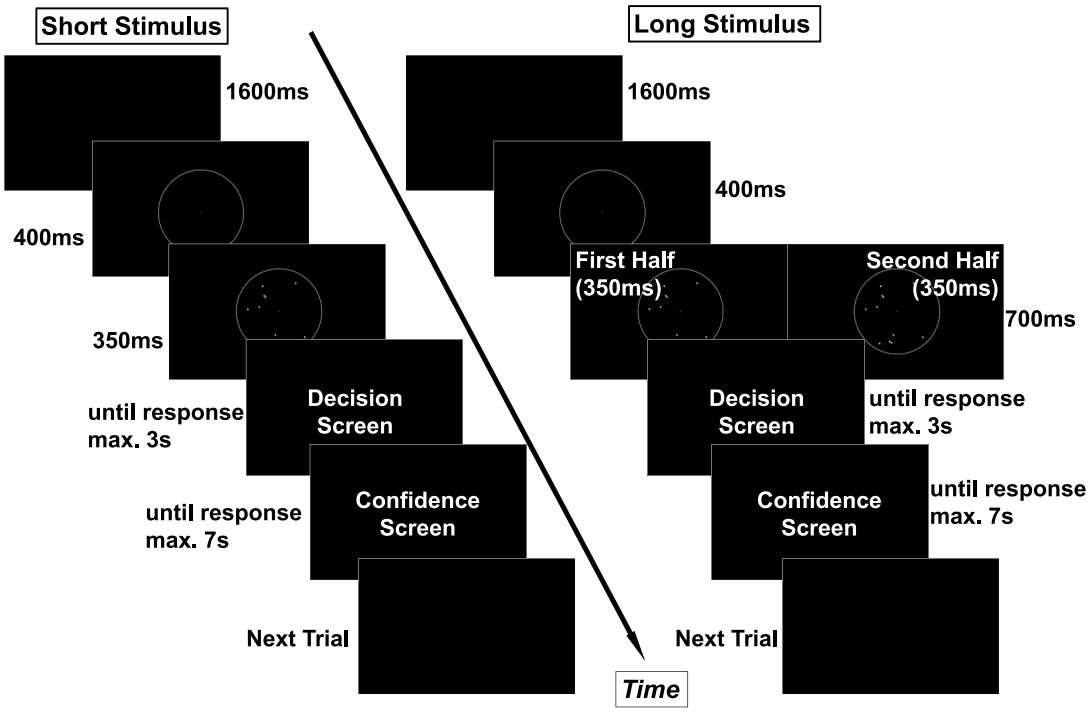


Fig. 10. Trial structure in Experiment 2.
Note. At the beginning of a trial, a blank screen was presented for 1600 ms, followed by an empty aperture with a fixation cross in the center which lasted 400 ms. After that, either a short (350 ms) or a long (700 ms) stimulus was shown. Then the decision screen and confidence screen were provided. The confidence screen can be seen in Fig. 11.

approved by the local ethics committee (reference number 2022_0127_246).

The experiment code was written in MATLAB using the Psychophysics Toolbox extensions (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997). The experimental MATLAB code, data, and analyses in R are freely available under osf.io/krw53. The experiment was preregistered at https://aspredicted.org/mvnr-w7m8.pdf.

Participants

Twelve cognitive science students (aged 20–26, 11 right-handed and 1 left-handed, 6 female and 6 male) from the University of Tübingen, who were naive to the purpose of the experiment, were recruited as participants. All participants had normal or corrected-to-normal vision and provided informed consent before the experiment. Participants received either course credits or monetary compensation (12.5 Euros per hour). We determined the number of participants, $n = 12$, by a power analysis ensuring that the expected performance difference between the low and high coherence conditions (accuracy 65% versus 71% accuracy, based on the data of Fleming et al., 2018) can be identified with $1 - \beta = 95\%$ power at a significance level of $\alpha = 5\%$.

Procedure

The general procedure and equipment were the same as in Experiment 1. However, this experiment consisted of only two sessions. In each session, participants performed two practice blocks followed by eight experimental blocks. Practice blocks consisted of four trials and each experimental block consisted of 80 trials, with each condition presented five times. Thus, participants performed $8 \cdot 80 = 640$ experimental trials per session for a total of $8 \cdot 80 \cdot 2 = 1280$ trials. The stimuli contained in each block were the same across participants, while the order of the trials within a block was randomized for each participant. Participants were encouraged to take self-timed breaks between blocks.

Each trial followed the same structure (Fig. 10): A blank screen was shown for 1600 ms followed by an empty aperture with a fixation cross lasting 400 ms. Participants then saw the RDK stimulus, made a binary decision about the motion direction (left or right), and provided a confidence rating for that decision. The stimulus durations of 350 ms in the short and 700 ms in the long condition were consistent with those in the two intervals in Rollwage et al. (2020).

Stimuli

We adopted the RDK stimulus display from Fleming et al. (2018). To stay consistent with their stimulus presentation on a monitor with a refresh rate of 60 Hz, we presented each stimulus frame twice on our 120 Hz monitor. In each RDK, white (103 cd/m^2) dots with a diameter of 0.12 deg were presented within a white circular aperture of 7 deg diameter on a black (0.08 cd/m^2) background. A total number of 30 dots moved at a speed of 5 deg/s. In a low-coherence condition, two dots (a proportion of 6.7%) moved horizontally either left or right. In a high-coherence condition, it was four dots (13.3%). As an exception, in the first practice block, the coherence was set to an extremely high value of 93%, to make participants familiar with the trial procedure. In the short condition (350 ms), we presented either low or high coherence motion. This resulted in four conditions for the short presentation duration corresponding to the cross combinations {left, right} × {low, high}.

In the long condition (700 ms, matching that of Rollwage et al., 2020), we again presented the four conditions in the first interval but now paired with either of three second-interval conditions: In the second interval, we presented either congruent, incongruent, or ambiguous motion. In the congruent condition, the coherent motion in the second interval was in the same direction as the first: two out of 30 (6.7%) dots moved coherently in the same direction. In the incongruent condition, the direction reversed and two out of 30 (6.7%) dots moved coherently in the opposite direction. In the ambiguous condition, two dots moved in the same direction as in the first interval and two other dots moved in the opposite direction. Thus, we had in total 16 conditions: Four short conditions corresponding to the cross-combinations of {left, right}

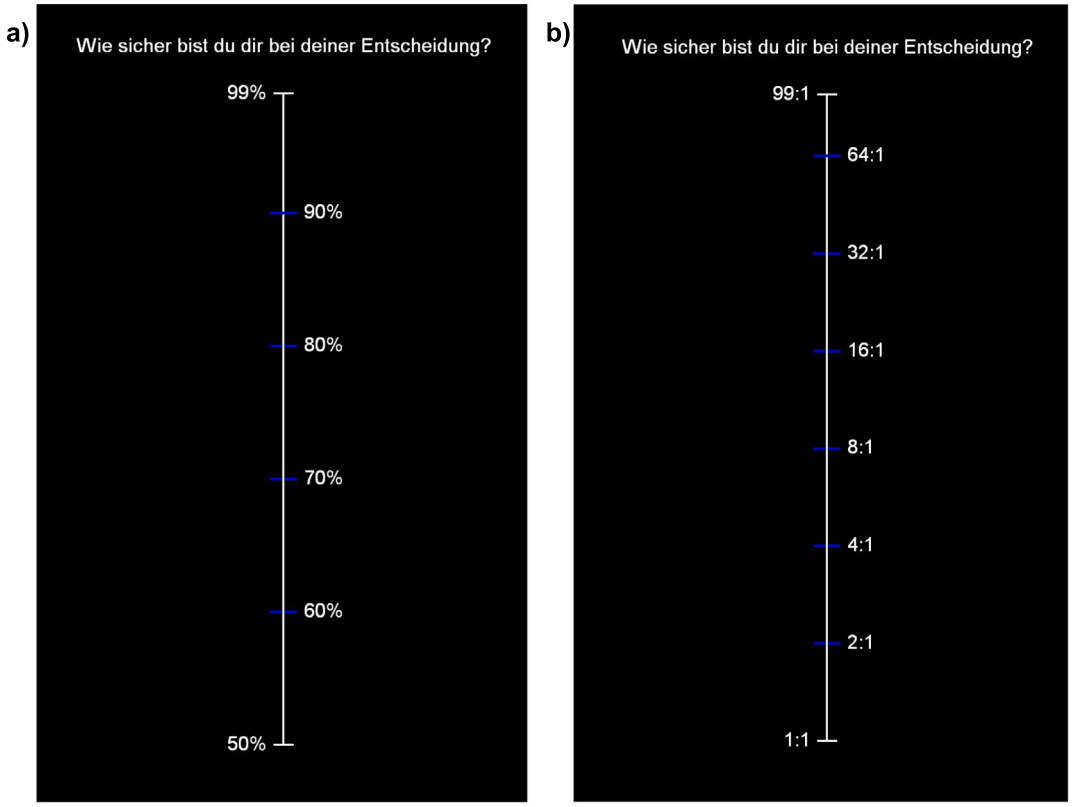


Fig. 11. Probability and logarithmic odds visual analogue scales for confidence ratings in Experiment 2.
Note. Participants gave confidence ratings in one session as (a) linearly scaled probabilities between 50% and 99% and in another session as (b) logarithmically scaled odds between 1 : 1 and 99 : 1. For visualization purposes, we added blue marks on both scales in this figure to demonstrate the scaling but these were not shown to participants. In the experiment, participants only saw one mark indicating the current position of the cursor with the corresponding value on the right. By mouse click, participants confirmed their choice.

× {low, high} and 12 long conditions corresponding to {left, right} × {low, high} × {congruent, incongruent, ambiguous}.

Similar to Experiment 1, we reduced noise due to random sampling of stimuli in the different conditions (see, Barthelmé & Mamassian, 2009) by generating stimuli in the short condition and deriving matched stimuli in the other conditions. As a basis, we generated short, rightward motion with low and high coherence. For the leftward motion stimuli, dot motions were mirrored in the middle of the screen. For the second interval, we paired this short stimulus motion with another short stimulus motion and concatenated both. This way, over the experiment but in different trials, the dot motion in the second interval exactly matched that of the first interval. These matched stimuli were then used for both sessions and all participants in randomized order.

For such stimuli and exactly this stimulus presentation duration, Zylberberg et al. (2012) found that decisions mostly depended on the dot motion at around 200 ms. Later motion energy had a decaying impact on the decision. Confidence formation depended almost exclusively on the stimulus motion supporting the decision but not on the stimulus motion contrasting it. This finding was then extended by Rollwage et al. (2020) who found that higher confidence facilitated the accumulation of evidence in favor of the previously made decision and reduced the probability of the change-of-mind behavior. From both findings, we predicted that evidence accumulation in our ambiguous condition would be mostly similar to that in the congruent condition because incongruent stimulus motion seems to be mostly disregarded in this task.

After the RDK was shown, the decision screen appeared and participants had to indicate by mouse click whether the motion direction was, on average, to the left or to the right. Participants had to respond within 3 s, otherwise the decision screen was replaced by a request to respond faster. In these cases, the trial with the presented stimulus was repeated at a random time in the remainder of the block. If participants responded in time, the trial continued.

Confidence ratings

Next, participants rated their confidence in this decision on a vertical visual analogue scale as shown in Fig. 11. We manipulated this confidence scale between the sessions in a balanced way. Half of the participants first used a scale that linearly mapped the position on the scale to probabilities ranging from 50% to 99%, see Fig. 11a. These participants were then presented with a logarithmic odds scale in the second session: Instead of probabilities, odds were presented ranging from 1 : 1 to 99 : 1 and these odds were spaced logarithmically across the scale, Fig. 11b. This scale was explained to the participants in the practice blocks: For example, an odds-value of 2 : 1 corresponds to the probability of 67% and indicated that, in trials with this confidence rating, for every incorrect response there should be two correct responses. The odds values were rounded. This scale was inspired by Phillips and Edwards (1966), who found better calibration of participants' raw confidence ratings when using logarithmic scaling. To assess whether calibration did indeed differ between the two conditions, we calculated the expected calibration error (ECE; Guo et al., 2017; as in Experiment 1 where we used it for calculating incentives). However, we removed the monetary incentive with the goal of removing overconfidence bias in raw confidence ratings (Krawczyk, 2012).

3.2. Results

We again computed correlations between accuracy, decision time, (raw) confidence ratings, and confidence decision time as a validation

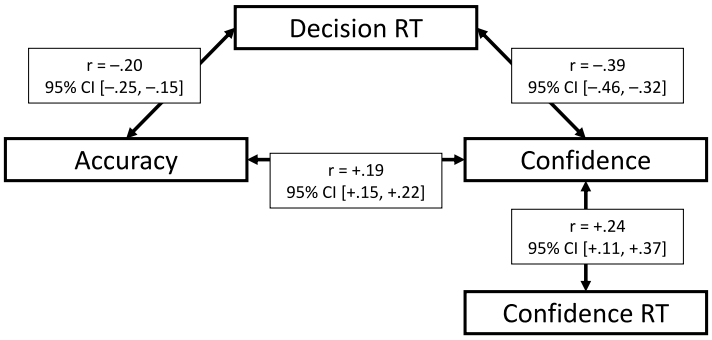


Fig. 12. Correlation of main variables in Experiment 2.
Note. As in Fig. 6 but for Experiment 2, for each participant, a correlation was computed between the accuracy (percentage of correct responses over all conditions), decision response times (RT), confidence ratings, and response times in giving these confidence ratings. These individual correlations were Fisher z-transformed to compute an average and confidence intervals, which were then transformed back to the correlation values shown here.

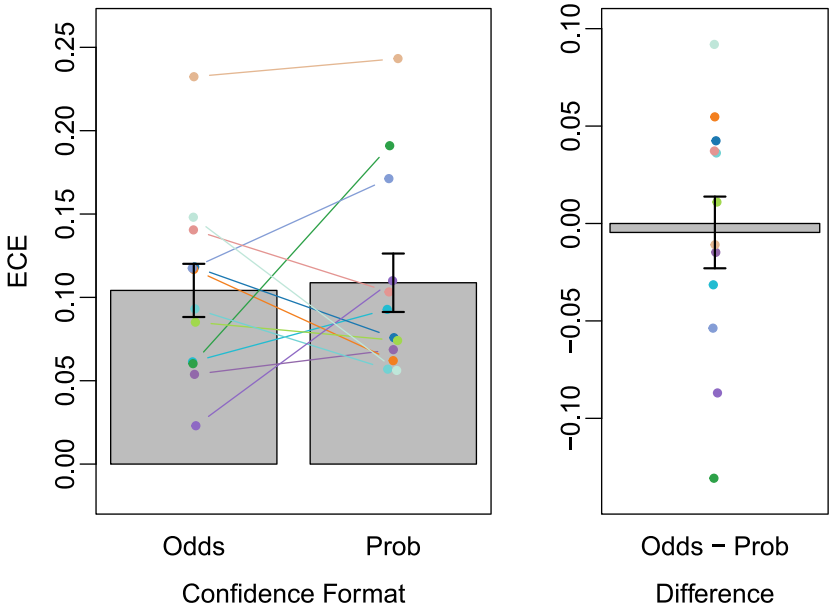


Fig. 13. Comparison between confidence rating scales.
Note. Contrasting participants' raw confidence rating calibration using the two scales shown in Fig. 11, we found no difference between them. Error bars indicate standard errors.

for the experimental manipulation (Fig. 12). In line with Experiment 1 and the typical findings (Lee et al., 2023; Moran et al., 2015; Pleskac & Busemeyer, 2010), accurate decisions were made faster than inaccurate decisions. This correlation suggests slow errors as found in previous work on RDKs (Forstmann et al., 2016; Lee et al., 2023; Ratcliff et al., 2016). Consistent with that, confidence ratings also correlated positively with accuracy and negatively with decision time. Again, the latter correlation was stronger than Rahnev et al. (2020) found on average across many studies. By and large, this correlation pattern can be seen as a validation of our experiment. However, in contrast to Experiment 1 and unexpected to us, we found a positive rather than a close-to-zero correlation between confidence and confidence response time.

Calibration

To evaluate calibration of participants' raw confidence ratings, we compared ECE when participants used the linear probability scale (average ECE = 0.11, 95% CI $[0.07, 0.15]$) to that when they used the logarithmic odds scale (average ECE = 0.10, 95% CI $[0.07, 0.14]$). Contrary to our expectation, we found no advantage for either scale, average difference in ECE = 0.00, 95% CI $[-0.05, 0.04]$, Fig. 13.

We then calibrated participants' confidence ratings as before. Calibration was again done separately for each participant, condition, and session to avoid confounds. As shown in Fig. 14, the raw confidence ratings deviated from perfect calibration as before. But in contrast to Experiment 1, participants were not overconfident but underconfident before calibration (perhaps explained by the missing incentive in Experiment 2, Krawczyk, 2012) with an average ECE = 0.10, 95% CI $[0.07, 0.13]$. A five-fold cross-validation with 10 repeats showed a good alignment of confidence ratings with accuracies after the calibration, average ECE = 0.03, 95% CI $[0.02, 0.03]$. Comparing ECE before and after calibration yielded a clear improvement of 0.08 ECE-points, 95% CI $[0.05, 0.11]$.

Evidence

Evidence accumulation trajectories for the different conditions are shown in Fig. 15. Importantly, we computed evidence accumulation in the direction of the initially presented motion so that incongruent information corresponds to a decrease in evidence by definition. In the short presentation, the low coherence condition ($\bar{E}^{\text{low}}_{[0,a]} = 0.65$, 95% CI $[0.32, 0.99]$) allowed for less evidence accumulation than the

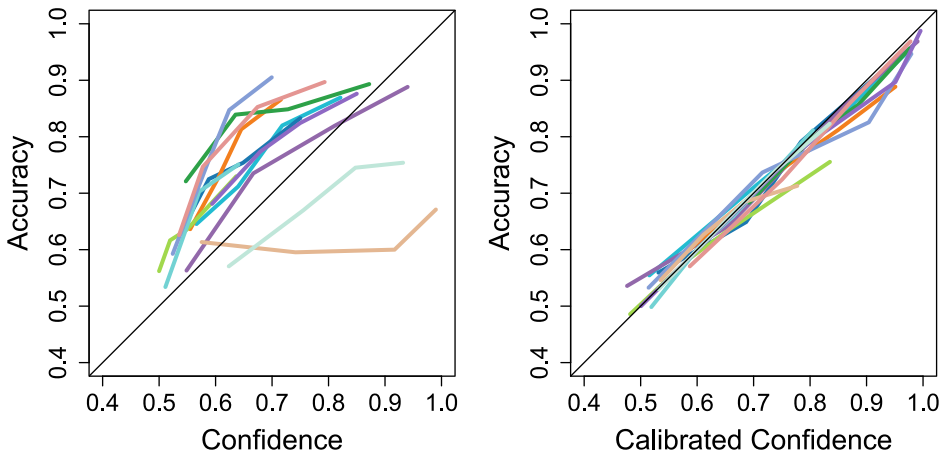


Fig. 14. Confidence calibration in Experiment 2.
Note. Similar to Fig. 7 but for Experiment 2, we show the relation of confidence ratings (x-axis) and accuracy (y-axis) before (left) and after calibration (right) with one line per participant. Diagonal lines represent perfect calibration.

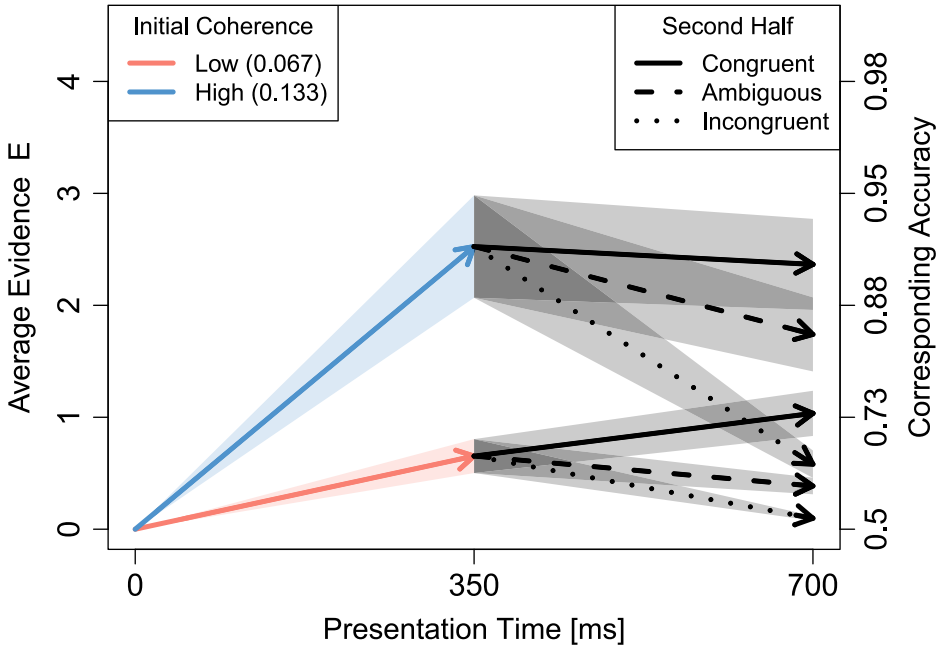


Fig. 15. Dependencies in evidence accumulation across time.
Note. Evidence trajectories across stimulus presentation time: High coherence led to more evidence accumulation in the first interval. In the second interval, the accumulation continues depending on the congruency of the second interval. With high coherence in the first interval, incongruent evidence had a stronger, negative impact on accumulated evidence. Shaded bands around trajectories represent standard errors interpolated between the time steps.

high coherence condition ($\bar{E}^{\text{high}}_{[0,a]} = 2.52$, 95% CI $[1.52, 3.53]$) with a difference of $\bar{E}^{\text{high}}_{[0,a]} - \bar{E}^{\text{low}}_{[0,a]} = 1.87$ (95% CI $[1.12, 2.62]$).

Evidence accumulation from congruent motion in the second interval depended on the first interval of stimulus presentation. When little evidence was accumulated in the first interval (due to low coherence), the second interval contributed a substantial amount of evidence in addition to that of the first interval, $\bar{E}^{\text{low,congruent}}_{[a,b]} = 0.38$, 95% CI $[0.17, 0.59]$. This is shown in Fig. 15 as the solid line continuing the trajectory of the lower, red line. In contrast, when the first interval provided more evidence (due to high coherence), participants did not accumulate further evidence in the second interval, $\bar{E}^{\text{high,congruent}}_{[a,b]} = -0.16$, 95% CI $[-0.51, 0.19]$. This is a substantial difference in evidence accumulation in the second interval depending on the first, ($\bar{E}^{\text{high,congruent}}_{[a,b]} - \bar{E}^{\text{low,congruent}}_{[a,b]} = -0.54$, 95% CI $[-0.96, -0.13]$).

A similar dependence was present when participants were shown incongruent motion in the second interval of stimulus presentation. Because the incongruent motion also had a low coherence, we found that this almost canceled the evidence accumulated in the low coherence first interval: The decrease in evidence from the incongruent second interval ($\bar{E}^{\text{low,incongruent}}_{[a,b]} = -0.56$, 95% CI $[-0.88, -0.23]$) substantially reduced the initially accumulated evidence. Nevertheless, we again found a reduced evidence accumulation in the second half (which was now negative) so that some evidence remained towards the motion direction of the first interval, $\bar{E}^{\text{low, incongruent}}_{[0,b]} = 0.10$, 95% CI $[0.05, 0.14]$. In principle, this could be seen as a primacy effect in which the initial part of the stimulus presentation is given more weight—for which we here provide a well-interpretable numerical quantification.

When participants saw high-coherence motion in the first interval, incongruent motion in the second interval evoked a stronger

reduction in evidence ($\bar{E}^{\text{high, incongruent}}_{[a,b]} - \bar{E}^{\text{low, incongruent}}_{[a,b]} = -1.39$, 95% CI $[-2.06, -0.72]$): Given the same low-coherent, incongruent motion strength, participants revised their judgments much more when previous evidence was high rather than when it was low. As a result, when the first interval featured a high coherence, participants accumulated evidence at the end of the incongruent condition to a degree ($\bar{E}^{\text{high, incongruent}}_{[0,b]} = 0.58$, 95% CI $[0.31, 0.85]$) which was comparable to that in the low-coherence, short presentation duration.

The ambiguous condition had an effect that is more comparable with that of the incongruent condition: Although the second interval partially showed evidence congruent with the first interval, the evidence accumulated in the first interval tended to be reduced by the second interval ($\bar{E}^{\text{low,ambiguous}}_{[a,b]} = -0.27$, 95% CI $[-0.59, 0.06]$ and $\bar{E}^{\text{high,ambiguous}}_{[a,b]} = -0.78$, 95% CI $[-1.30, -0.27]$). This reduction depended on how much evidence participants extracted from the first interval: High coherence in the first interval was followed by a greater reduction in evidence compared to a low coherence ($\bar{E}^{\text{high, ambiguous}}_{[a,b]} - \bar{E}^{\text{low, ambiguous}}_{[a,b]} = -0.52$, 95% CI $[-1.05, 0.02]$). Thus, incongruent evidence in the second interval of stimulus presentation seemed to be weighted more than congruent evidence.

3.3. Discussion

In Experiment 2, we investigated how accumulated evidence during the second presentation interval depends on the first. For all conditions, the evidence accumulated in the second interval was modulated by the evidence accumulated in the first interval. In contrast to low coherence in the first interval, high coherence leads to more initial evidence accumulation resulting in less evidence accumulation (or more contrary evidence accumulation) in the second interval.

In part, this is consistent with previous results where also less evidence is accumulated in the second interval (Fleming et al., 2018; Rollwage et al., 2020; Zylberberg et al., 2012), where the decision formation primarily depended on the first stimulus presentation interval. Even in the low-coherence conditions, we found a decelerating accumulation trend: Low-coherence in the first interval produces an average evidence of $\bar{E}^{\text{low}}_{[0,a]} = 0.65$ and continued presentation of the same strength of coherent motion only yielded additional evidence of $\bar{E}^{\text{low,congruent}}_{[a,b]} = 0.38$. This is also in line with the idea that participants accumulate evidence up to a bound of internal certainty (a desired level of confidence, Hausmann & Läge, 2008). However, we did not find that the ambiguous condition produced similar results as the congruent condition.

Counter to what we expected, incongruent motion had a larger impact when participants had initially accrued more evidence in the high-coherence condition than in the low-coherence condition. Low-coherence, incongruent motion in the second interval counteracted the low-coherence evidence from the first interval producing average evidence close to 0. But if the first interval had a high coherence, subsequently presenting low-coherence, incongruent motion evoked a strong reduction in evidence and, in total, yielded roughly the same amount of evidence ($\bar{E}^{\text{high, incongruent}}_{[0,b]} = 0.58$) as only showing low-coherence evidence in the first interval ($\bar{E}^{\text{low}}_{[0,a]} = 0.65$).

Note that these results are separate from (dis-)confirmation bias, which pertains to differential effects of congruent and incongruent evidence on decision confidence (Boldt & Desender, 2023; Peters et al., 2017; Rollwage et al., 2020; Zylberberg et al., 2012). Our results are however not in line with studies demonstrating a confirmation bias on decisions themselves (Glickman et al., 2022; Peters, 2022; Talluri et al., 2018). We elaborate on the role of biases on confidence ratings later.

Also contrary to our expectations, the choice of confidence rating scale did not seem to change calibration. It is possible that the advantage of the logarithmic scales in Phillips and Edwards (1966) arose from their experimental design which was remotely similar to our toy example with biased coin flips. A perceptual task such as the one we presented here may not produce the same effect. Different stimulus materials may fit particular confidence rating scales but we are unaware of a systematic investigation of this.

4. General discussion

Based on the mathematical principles of SPRT (on which DDMs are also based), we suggest to directly measure participants' accumulated evidence after viewing stimuli for different presentation durations. Our approach uses participants' decisions and confidence ratings and transforms them into evidence as the logarithmic odds of calibrated confidence ratings in each trial. The main advantage of this approach lies in the additivity: Evidence accumulated from a short ($\bar{E}_{[0,a]}$) and long first presentation interval ($\bar{E}_{[0,b]}$) can be easily algebraically decomposed to yield the evidence that was accumulated during the second interval of stimulus presentation, $\bar{E}_{[a,b]} = \bar{E}_{[0,b]} - \bar{E}_{[0,a]}$. This allows for a direct comparison between the amount of evidence accumulated in the first versus second interval, $\bar{E}_{[0,a]}$ versus $\bar{E}_{[a,b]}$.

Applying this method in two experiments, we have demonstrated decelerating evidence accumulation in two different dot motion stimuli. This conceptually replicates the results of previous studies indicating that a decision is made based on the early phases of stimulus presentation (Fleming et al., 2018; Rollwage et al., 2020; Zylberberg et al., 2012). Moreover, we have gathered tentative evidence for an effect of motion-causality on the evidence accumulation rate in Experiment 1. Experiment 2 showed that in continuous presentation of the stimuli and when participants did not give an explicit response in between the stimulus intervals, the impact of incongruent motion was increased when initial evidence was high.

4.1. Improving theory testing through additive quantification of evidence

Measuring accumulated evidence in the way we suggest here allows researchers to better interpret effect sizes. As already discussed in Experiment 1, there was an advantage for evidence accumulation when presenting causal rather than anti-causal motion for short ($\bar{E}^{\text{causal}}_{[0,a]} = 0.55$ versus $\bar{E}^{\text{anti-causal}}_{[0,a]} = 0.41$) as well as for long durations ($\bar{E}^{\text{causal}}_{[0,b]} = 0.93$ versus $\bar{E}^{\text{anti-causal}}_{[0,b]} = 0.79$). In both, short and long, this difference was constant at $\bar{E}^{\text{causal}} - \bar{E}^{\text{anti-causal}} = 0.14$. Taking this result at face value hints towards the development of an advantage from causal displays in the early stages of perception which does not compound in the later stages. Such numerical evaluations are only possible due to the additive quantification of evidence. Other measures—such as accuracy—do not allow such interpretations (at least not without making further assumptions on the underlying noise distribution). However, for now, we want to caution about interpreting this result too hastily: Whether the effect is truly due to the causality manipulation or due to confounded statistical properties of the particular stimulus presentation that we chose remains to be shown. Conceptual replications are necessary to corroborate this result.

In general, this quantification of accumulated evidence allows researchers to go beyond the traditional hypothesis testing approach (be it Frequentist or Bayesian) and towards proper quantification of effects (Calin-Jageman & Cumming, 2019; Cumming, 2013; Kruschke & Liddell, 2017). This avoids just predicting the direction of an effect but leaving the effect sizes underspecified, which is a problem in many fields of psychological research (Dienes, 2008; Meehl, 1967).

4.2. Interpretability issues due to biased confidence ratings and suboptimal integration

The price researchers have to pay for the intriguing properties of our evidence measure lies in the reliance on confidence ratings. Confidence ratings are affected by various factors (Kiani et al., 2014), primarily, when these factors are informative about general visibility (Hellmann et al., 2024; Rausch et al., 2018; Rausch & Zehetleitner, 2019). For example, participants in our experiment may have taken the presentation duration of a stimulus as an indicator of uncertainty and provided higher confidence ratings for long and lower for short presentation durations. While stimulus duration is a valid cue in the

real world, our Experiment 2 deliberately disentangled stimulus duration from evidence. This turns participants' heuristic into a bias. There are many more factors that can produce similar biases (Fleming, 2024; Hellmann et al., 2023, 2024; Moran et al., 2015; Peters, 2022; Pleskac & Busemeyer, 2010; Rahnev & Denison, 2018; Sánchez-Fuenzalida et al., 2025). It is also noteworthy that simply asking participants to provide confidence responses itself may have an impact on how the stimuli are processed (Dou et al., 2024).

Given that many factors can influence confidence ratings, our accumulated evidence measure—being based on confidence ratings—should not be interpreted as an exclusive measure of sensory evidence. Instead, it is best understood as a performance measure with the intriguing additivity property. Biases that make participants overestimate their confidence in some trials and underestimate it in others will reduce this measure. This reduction is similar to the reduction due to noise in confidence ratings (c.f. Figs. 3 and 4). This is what a measure of evidence—understood as a performance measure—ought to do: Less informative confidence ratings, be it due to bias or noise, are appropriately reflected in reduced amounts of evidence.

To increase interpretability and eliminate the effect of some of these biases, calibration should therefore be done separately for each condition: When participants overestimate confidence in the long stimulus presentation condition and underestimate confidence in the short, separate calibration in these conditions avoids an artificial reduction of evidence. But this, in turn, comes at the cost of requiring many trials to guarantee sufficiently reliable calibration within each condition. In our experiments and throughout multiple sessions, each participant gave responses throughout 768 and 1280 trials in Experiment 1 and 2, respectively. This positions our approach among the psychophysical tradition of studying few participants exhaustively (Normand, 2016; Rouder & Haaf, 2018; Smith & Little, 2018), which is also a requirement when using DDMs to achieve sufficiently good model fits.

Finally, we want to repeat that accumulated evidence can be reduced by a reduction in discrimination performance, a reduction in informativeness of confidence ratings, or a mixture of both. These two components, called Type I and Type II performance in metacognitive research (Clarke et al., 1959; Fleming & Lau, 2014; Peters, 2022), are not differentiated by our measure.

Why then use this measure of accumulated evidence when its interpretability is somewhat limited? In short, because the traditional alternatives have the same downsides without the upside of additivity. Accuracy does not consider confidence ratings at all and d' makes a normal-distribution assumption instead of measuring confidence. But by the theoretical principles of SPRT, uncertainty evaluation is an integral part of evidence and, therefore, should be measured. Moreover, when combining information from two stimulus intervals, accuracy or d' also suffer from the ambiguity of potentially suboptimal integration of information.

4.3. Accumulated evidence: Overt performance measure versus internal variable

Crucially, the measure of accumulated evidence we suggest here should be considered as a performance metric on overt responses. It is agnostic about when and how evidence is processed internally. This is in stark contrast to DDMs, which estimate the parameters governing evidence accumulation and consider evidence as a latent variable. Starting from this perspective, multiple studies have found relations of the latent evidence variable to neural activity (e.g., in the lateral intraparietal area, LIP, Drugowitsch et al., 2012; Gold & Shadlen, 2007; Kiani & Shadlen, 2009; Yang & Shadlen, 2007; but also frontal areas, Fleming et al., 2018; and others Purcell et al., 2010). Ultimately, this perspective also promises to obtain explicit measures of evidence on a trial by trial basis using neural imaging techniques. But here, we took the complementary perspective. We measured evidence from behavioral responses. At the core, we combined (1) the mathematical principles also underlying DDMs in which evidence accumulation is reflected in summing up logarithmic likelihood ratios (Griffith et al., 2021; Jaynes, 2003; Lange et al., 2021) and (2) confidence ratings reflecting these likelihoods after calibration (Kepecs & Mainen, 2012; Meyniel et al., 2015; Pouget et al., 2016).

Both of these perspectives may ultimately converge into a unified framework explaining the two sides of evidence accumulation: the internal computational mechanisms and their translation into behaviorally measurable responses. Although coming from the model-agnostic perspective, the here presented approach would greatly benefit from developing generative models of confidence (Guggenmos, 2022), specifically those that relate confidence to the internally accumulated evidence (Balsdon et al., 2020; Calder-Travis et al., 2023; Lee et al., 2023; Pereira et al., 2022).

4.4. Conclusion

The application of methods from the sequential probability ratio tests allows quantifying the accumulated evidence in an additive way. This brings advantages to the comparability and interpretability of results, but also rests on the proper evaluation of participants' confidence ratings. Future applications of this method could convert the treatment of evidence from being a latent, internal variable into a performance measure, which allows for generating more specific hypotheses in psychophysical research.

CRediT authorship contribution statement

Sascha Meyen: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Lin Lin:** Writing – review & editing, Writing – original draft, Visualization, Validation, Investigation, Formal analysis, Data curation, Conceptualization. **Carina Schrenk:** Writing – review & editing, Visualization, Validation, Investigation, Formal analysis, Data curation, Conceptualization. **Volker H. Franz:** Writing – review & editing, Supervision, Software, Resources, Project administration, Funding acquisition, Conceptualization.

Funding

This work was supported by the German Research Foundation (DFG): SFB 1233, Robust Vision: Inference Principles and Neural Mechanisms, TP C1, project number: 276693517; the Institutional Strategy of University of Tübingen (DFG, ZUK 63); and the Cluster of Excellence "Machine Learning: New Perspectives for Science", EXC 2064/1, number 390727645.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Madeleine Soukup for her work during this project.

Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.newideapsych.2026.101249.

Data availability

The experimental MATLAB code, data, and analyses in R are freely available under osf.io/krw53.

References

Adler, W. T., & Ma, W. J. (2018). Comparing Bayesian and non-Bayesian accounts of human confidence reports. *PLoS Computational Biology*, *14*(11), Article e1006572. http://dx.doi.org/10.1371/journal.pcbi.1006572.

Aitchison, L., Bang, D., Bahrami, B., & Latham, P. E. (2015). Doubly Bayesian analysis of confidence in perceptual decision-making. *PLoS Computational Biology*, *11*(10), Article e1004519. http://dx.doi.org/10.1371/journal.pcbi.1004519.

Balsdon, T., Wyart, V., & Mamassian, P. (2020). Confidence controls perceptual evidence accumulation. *Nature Communications*, *11*(1), http://dx.doi.org/10.1038/s41467-020-15561-w.

Barthelmé, S., & Mamassian, P. (2009). Evaluation of objective uncertainty in the visual system. *PLoS Computational Biology*, *5*(9), Article e1000504. http://dx.doi.org/10.1371/journal.pcbi.1000504.

Bertana, A., Chetverikov, A., van Bergen, R. S., Ling, S., & Jehee, J. F. M. (2021). Dual strategies in human confidence judgments. *Journal of Vision*, *21*(5), 21. http://dx.doi.org/10.1167/jov.21.5.21.

Bitzer, S., Park, H., Blankenburg, F., & Kiebel, S. J. (2014). Perceptual decision making: drift-diffusion model is equivalent to a Bayesian model. *Frontiers in Human Neuroscience*, *8*, 102. http://dx.doi.org/10.3389/fnhum.2014.00102.

Bogacz, R. (2007). Optimal decision-making theories: linking neurobiology with behaviour. *Trends in Cognitive Sciences*, *11*(3), 118–125. http://dx.doi.org/10.1016/j.tics.2006.12.006.

Boldt, A., & Desender, K. (2023). Dis-confirmatory evidence drives confidence. In *2023 conference on cognitive computational neuroscience* (pp. 436–438). Cognitive Computational Neuroscience.

Boundy-Singer, Z. M., Ziemba, C. M., & Goris, R. L. (2023). Confidence reflects a noisy decision reliability estimate. *Nature Human Behaviour*, *7*(1), 142–154. http://dx.doi.org/10.1038/s41562-022-01464-x.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436, Retrieved from http://psychtoolbox.org/.

Burr, D., Banks, M. S., & Morrone, M. C. (2009). Auditory dominance over vision in the perception of interval duration. *Experimental Brain Research*, *198*(1), 49–57. http://dx.doi.org/10.1007/s00221-009-1933-z.

Calder-Travis, J., Bogacz, R., & Yeung, N. (2023). Expressions for Bayesian confidence of drift diffusion observers in fluctuating stimuli tasks. *Journal of Mathematical Psychology*, *117*, Article 102815. http://dx.doi.org/10.1016/j.jmp.2023.102815.

Calin-Jageman, R. J., & Cumming, G. (2019). The new statistics for better science: Ask how much, how uncertain, and what else is known. *The American Statistician*, *73*, 271–280. http://dx.doi.org/10.1080/00031305.2018.1518266.

Clarke, F. R., Birdsall, T. G., & Tanner Jr, W. P. (1959). Two types of ROC curves and definitions of parameters. *Journal of the Acoustical Society of America*, *31*(5), 629–630. http://dx.doi.org/10.1121/1.1907764.

Cumming, G. (2013). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. http://dx.doi.org/10.1177/0956797613504966.

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(4), 429–453. http://dx.doi.org/10.3758/cabn.8.4.429.

de Gardelle, V., & Summerfield, C. (2011). Robust averaging during perceptual judgment. *Proceedings of the National Academy of Sciences*, *108*(32), 13341–13346. http://dx.doi.org/10.1073/pnas.1104517108.

Desender, K., Donner, T. H., & Verguts, T. (2021). Dynamic expressions of confidence within an evidence accumulation framework. *Cognition*, *207*, Article 104522. http://dx.doi.org/10.1016/j.cognition.2020.104522.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Bloomsbury Publishing.

Dou, W., Afrakhteh, S., & Samaha, J. (2024). Metacognitive introspection alters the dynamics of pre-decisional neural evidence accumulation. http://dx.doi.org/10.1101/2024.11.26.625501, bioRxiv.

Dou, W., Martinez Arango, L. J., Castaneda, O. G., Arellano, L., Mcintyre, E., Yballa, C., & Samaha, J. (2024). Neural signatures of evidence accumulation encode subjective perceptual confidence independent of performance. *Psychological Science*, *35*(7), 760–779. http://dx.doi.org/10.1177/09567976241246561.

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., & Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *The Journal of Neuroscience*, *32*(11), 3612–3628. http://dx.doi.org/10.1523/jneurosci.4010-11.2012.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433. http://dx.doi.org/10.1038/415429a.

Fetsch, C. R., Pouget, A., DeAngelis, G. C., & Angelaki, D. E. (2011). Neural correlates of reliability-based cue weighting during multisensory integration. *Nature Neuroscience*, *15*(1), 146–154. http://dx.doi.org/10.1038/nn.2983.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika*, *10*(4), 507–521. http://dx.doi.org/10.2307/2331838.

Fleming, S. M. (2024). Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, *75*(1), 241–268. http://dx.doi.org/10.1146/annurev-psych-022423-032425.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443. http://dx.doi.org/10.3389/fnhum.2014.00443.

Fleming, S. M., van der Putten, E. J., & Daw, N. D. (2018). Neural mediators of changes of mind about perceptual decisions. *Nature Neuroscience*, *21*(4), 617–624. http://dx.doi.org/10.1038/s41593-018-0104-6.

Forstmann, B., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual Review of Psychology*, *67*(1), 641–666. http://dx.doi.org/10.1146/annurev-psych-122414-033645.

Franz, V. H. (2007). Ratios: A short guide to confidence limits and proper use. arXiv, Retrieved from https://arxiv.org/abs/0710.2024.

Gallagher, R. M., Suddendorf, T., & Arnold, D. H. (2019). Confidence as a diagnostic tool for perceptual aftereffects. *Scientific Reports*, *9*(1), 7124. http://dx.doi.org/10.1038/s41598-019-43170-1.

Gherman, S., & Philiastides, M. G. (2015). Neural representations of confidence emerge from the process of decision formation during perceptual choices. *NeuroImage*, *106*, 134–143. http://dx.doi.org/10.1016/j.neuroimage.2014.11.036.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. *Psychological Review*, *98*(4), 506–528. http://dx.doi.org/10.1037/0033-295X.98.4.506.

Glickman, M., Moran, R., & Usher, M. (2022). Evidence integration and decision confidence are modulated by stimulus consistency. *Nature Human Behaviour*, *6*(7), 988–999. http://dx.doi.org/10.1038/s41562-022-01318-6.

Gold, J. I., Law, C.-T., Connolly, P., & Bennur, S. (2008). The relative influences of priors and sensory evidence on an oculomotor decision variable during perceptual learning. *Journal of Neurophysiology*, *100*(5), 2653–2668. http://dx.doi.org/10.1152/jn.90629.2008.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*(1), 535–574. http://dx.doi.org/10.1146/annurev.neuro.29.051605.113038.

Green, D. M., & Swets, J. A. (1988). *Signal Detection Theory and Psychophysics*. Los Altos, CA: Peninsula.

Griffith, T., Baker, S.-A., & Lepora, N. F. (2021). The statistics of optimal decision making: Exploring the relationship between signal detection theory and sequential analysis. *Journal of Mathematical Psychology*, *103*, Article 102544. http://dx.doi.org/10.1016/j.jmp.2021.102544.

Guggenmos, M. (2022). Reverse engineering of metacognition. *ELife*, *11*, http://dx.doi.org/10.7554/eLife.75420.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning* (pp. 1321–1330). PMLR.

Hausmann, D., & Läge, D. (2008). Sequential evidence accumulation in decision making: The individual desired level of confidence can explain the extent of information acquisition. *Judgment and Decision Making*, *3*(3), 229–243. http://dx.doi.org/10.1017/s1930297500002436.

Hebart, M. N., Schriever, Y., Donner, T. H., & Haynes, J.-D. (2014). The relationship between perceptual decision variables and confidence in the human brain. *Cerebral Cortex*, *26*(1), 118–130. http://dx.doi.org/10.1093/cercor/bhu181.

Hellmann, S., Zehetleitner, M., & Rausch, M. (2023). Simultaneous modeling of choice, confidence, and response time in visual perception. *Psychological Review*, *130*(6), 1521–1543. http://dx.doi.org/10.1037/rev0000411.

Hellmann, S., Zehetleitner, M., & Rausch, M. (2024). Confidence is influenced by evidence accumulation time in dynamical decision models. *Computational Brain & Behavior*, *7*(3), 287–313. http://dx.doi.org/10.1007/s42113-024-00205-9.

Hochstein, S., & Ahissar, M. (2002). View from the top. *Neuron*, *36*(5), 791–804. http://dx.doi.org/10.1016/s0896-6273(02)01091-7.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Katyal, S., & Fleming, S. M. (2024). The future of metacognition research: Balancing construct breadth with measurement rigor. *Cortex*, *171*, 223–234. http://dx.doi.org/10.1016/j.cortex.2023.11.002.

Kauffmann, L., Chauvin, A., Guyader, N., & Peyrin, C. (2015). Rapid scene categorization: Role of spatial frequency order, accumulation mode and luminance contrast. *Vision Research*, *107*, 49–57. http://dx.doi.org/10.1016/j.visres.2014.11.013.

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *367*(1594), 1322–1337. http://dx.doi.org/10.1098/rstb.2012.0037.

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron*, *84*(6), 1329–1342. http://dx.doi.org/10.1016/j.neuron.2014.12.015.

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, *324*(5928), 759–764. http://dx.doi.org/10.1126/science.1169405.

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in psychtoolbox-3? *Perception*, *36*(14), 1–16, Retrieved from http://psychtoolbox.org/.

Koriat, A. (2011). Subjective confidence in perceptual judgments: A test of the self-consistency model. *Journal of Experimental Psychology: General*, *140*(1), 117–139. http://dx.doi.org/10.1037/a0022171.

Krawczyk, M. (2012). Incentives and timing in relative performance judgments: A field experiment. *Journal of Economic Psychology*, *33*(6), 1240–1246. http://dx.doi.org/10.1016/j.joep.2012.09.006.

Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. http://dx.doi.org/10.3758/s13423-016-1221-4.

Kullback, S. (1959). *Information Theory and Statistics*. New York: Wiley.

Laming, D. R. J. (1968). *Information Theory of Choice-Reaction Times*. Academic Press.

Lange, R. D., Chattoraj, A., Beck, J. M., Yates, J. L., & Haefner, R. M. (2021). A confirmation bias in perceptual decision-making due to hierarchical approximate inference. In M. A. K. Peters (Ed.), *PLoS Computational Biology*, *17*(11), Article e1009517. http://dx.doi.org/10.1371/journal.pcbi.1009517.

Lee, D. G., Daunizeau, J., & Pezzulo, G. (2023). Evidence or confidence: What is really monitored during a decision? *Psychonomic Bulletin & Review*, *30*(4), 1360–1379. http://dx.doi.org/10.3758/s13423-023-02255-9.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438. http://dx.doi.org/10.1038/nn1790.

Maniscalco, B., Charles, L., & Peters, M. A. K. (2025). Optimal metacognitive decision strategies in signal detection theory. *Psychonomic Bulletin & Review*, *32*(3), 1041–1069. http://dx.doi.org/10.3758/s13423-024-02510-7.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430. http://dx.doi.org/10.1016/j.concog.2011.09.021.

Maniscalco, B., & Lau, H. (2014). 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* (pp. 25–66). Springer, http://dx.doi.org/10.1007/978-3-642-45190-4_3.

Matejka, J., Glueck, M., Grossman, T., & Fitzmaurice, G. (2016). The effect of visual appearance on the performance of continuous sliders and visual analogue scales. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 5421–5432). ACM, http://dx.doi.org/10.1145/2858036.2858063.

Meding, K., Janzing, D., Schölkopf, B., & Wichmann, F. A. (2019). Perceiving the arrow of time in autoregressive motion. *Advances in Neural Information Processing Systems*, *32*.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*(2), 103–115.

Meyniel, F. (2020). Brain dynamics for confidence-weighted learning. *PLoS Computational Biology*, *16*(6), Article e1007935. http://dx.doi.org/10.1371/journal.pcbi.1007935.

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian probability: From neural origins to behavior. *Neuron*, *88*(1), 78–92. http://dx.doi.org/10.1016/j.neuron.2015.09.039.

Michel, M. (2022). Confidence in consciousness research. *WIREs Cognitive Science*, *14*(2), http://dx.doi.org/10.1002/wcs.1628.

Miyoshi, K., Webb, T., Rahnev, D., & Lau, H. (2025). Confidence and metacognition. *Encyclopedia of the Human Brain*, 252–268. http://dx.doi.org/10.31234/osf.io/6gyjf_v1.

Moran, R. (2015). Optimal decision making in heterogeneous and biased environments. *Psychonomic Bulletin & Review*, *22*(1), 38–53. http://dx.doi.org/10.3758/s13423-014-0669-3.

Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology*, *78*, 99–147. http://dx.doi.org/10.1016/j.cogpsych.2015.01.002.

Murphy, P. R., Robertson, I. H., Harty, S., & O'Connell, R. G. (2015). Neural evidence accumulation persists after choice to inform metacognitive judgments. *ELife*, *4*, http://dx.doi.org/10.7554/elife.11946.

Navajas, J., Bahrami, B., & Latham, P. E. (2016). Post-decisional accounts of biases in confidence. *Current Opinion in Behavioral Sciences*, *11*, 55–60. http://dx.doi.org/10.1016/j.cobeha.2016.05.005.

Newsome, W. T., Britten, K. H., & Movshon, J. A. (1989). Neuronal correlates of a perceptual decision. *Nature*, *341*(6237), 52–54. http://dx.doi.org/10.1038/341052a0.

Newsome, W. T., & Pare, E. B. (1988). A selective impairment of motion perception following lesions of the middle temporal visual area (MT). *Journal of Neuroscience*, *8*(6), 2201–2211. http://dx.doi.org/10.1523/JNEUROSCI.08-06-02201.1988.

Normand, M. P. (2016). Less is more: Psychologists can learn more by studying fewer people. *Frontiers in Psychology*, *7*, http://dx.doi.org/10.3389/fpsyg.2016.00934.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442, Retrieved from http://psychtoolbox.org/.

Pereira, M., Perrin, D., & Faivre, N. (2022). A leaky evidence accumulation process for perceptual experience. *Trends in Cognitive Sciences*, *26*(6), 451–461. http://dx.doi.org/10.1016/j.tics.2022.03.003.

Peters, M. A. (2022). Confidence in decision-making. In *Oxford research encyclopedia of neuroscience*. http://dx.doi.org/10.1093/acrefore/9780190264086.013.371.

Peters, M. A., Thesen, T., Ko, Y. D., Maniscalco, B., Carlson, C., Davidson, M., Doyle, W., Kuzniecky, R., Devinsky, O., Halgren, E., et al. (2017). Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature Human Behaviour*, *1*(7), 0139. http://dx.doi.org/10.1038/s41562-017-0139.

Philiastides, M. G., Ratcliff, R., & Sajda, P. (2006). Neural representation of task difficulty and decision making during perceptual categorization: A timing diagram. *The Journal of Neuroscience*, *26*(35), 8965–8975. http://dx.doi.org/10.1523/jneurosci.1655-06.2006.

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, *72*(3), 346, https://psycnet.apa.org/doi/10.1037/h0023653.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901. http://dx.doi.org/10.1037/a0019737.

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366–374.

Purcell, B. A., Heitz, R. P., Cohen, J. Y., Schall, J. D., Logan, G. D., & Palmeri, T. J. (2010). Neurally constrained modeling of perceptual decision making. *Psychological Review*, *117*(4), 1113–1143. http://dx.doi.org/10.1037/a0020311.

Rahnev, D., Balsdon, T., Charles, L., de Gardelle, V., Denison, R., Desender, K., Faivre, N., Filevich, E., Fleming, S. M., Jehee, J., Lau, H., Lee, A. L. F., Locke, S. M., Mamassian, P., Odegaard, B., Peters, M., Reyes, G., Rouault, M., Sackur, J., ... Zylberberg, A. (2022). Consensus goals in the field of visual metacognition. *Perspectives on Psychological Science*, *17*(6), 1746–1765. http://dx.doi.org/10.1177/17456916221075615.

Rahnev, D., & Denison, R. N. (2018). Suboptimality in perceptual decision making. *Behavioral and Brain Sciences*, *41*, Article e223. http://dx.doi.org/10.1017/S0140525X18000936.

Rahnev, D., Desender, K., Lee, A. L. F., Adler, W. T., Aguilar-Lleyda, D., Akdoğan, B., Arbuzova, P., Atlas, L. Y., Balcı, F., Bang, J. W., Bègue, I., Birney, D. P., Brady, T. F., Calder-Travis, J., Chetverikov, A., Clark, T. K., Davranche, K., Denison, R. N., Dildine, T. C., ... Zylberberg, A. (2020). The confidence database. *Nature Human Behaviour*, *4*(3), 317–325. http://dx.doi.org/10.1038/s41562-019-0813-1.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. http://dx.doi.org/10.1038/4580.

Rasanan, A. H. H., Rad, J. A., & Sewell, D. K. (2023). Are there jumps in evidence accumulation, and what, if anything, do they reflect psychologically? An analysis of Lévy flights models of decision-making. *Psychonomic Bulletin & Review*, *31*(1), 32–48. http://dx.doi.org/10.3758/s13423-023-02284-4.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. http://dx.doi.org/10.1037/0033-295x.85.2.59.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. http://dx.doi.org/10.1162/neco.2008.12-06-420.

Ratcliff, R., & Rouder, J. N. (2000). A diffusion model account of masking in two-choice letter identification. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(1), 127–140. http://dx.doi.org/10.1037/0096-1523.26.1.127.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281. http://dx.doi.org/10.1016/j.tics.2016.01.007.

Rausch, M., Hellmann, S., & Zehetleitner, M. (2018). Confidence in masked orientation judgments is informed by both evidence and visibility. *Attention, Perception, & Psychophysics*, *80*(1), 134–154. http://dx.doi.org/10.3758/s13414-017-1431-5.

Rausch, M., Hellmann, S., & Zehetleitner, M. (2023). Measures of metacognitive efficiency across cognitive models of decision confidence. *Psychological Methods*, http://dx.doi.org/10.1037/met0000634.

Rausch, M., & Zehetleitner, M. (2019). The folded X-pattern is not necessarily a statistical signature of decision confidence. *PLoS Computational Biology*, *15*(10), Article e1007456. http://dx.doi.org/10.1371/journal.pcbi.1007456.

Richter, T., Ulrich, R., & Janczyk, M. (2023). Diffusion models with time-dependent parameters: An analysis of computational effort and accuracy of different numerical methods. *Journal of Mathematical Psychology*, *114*, Article 102756. http://dx.doi.org/10.1016/j.jmp.2023.102756.

Rollwage, M., Loosen, A., Hauser, T. U., Moran, R., Dolan, R. J., & Fleming, S. M. (2020). Confidence drives a neural confirmation bias. *Nature Communications*, *11*(1), 2634. http://dx.doi.org/10.1038/s41467-020-16278-6.

Rouder, J. N., & Haaf, J. M. (2018). Power, dominance, and constraint: A note on the appeal of different design traditions. *Advances in Methods and Practices in Psychological Science*, *1*(1), 19–26. http://dx.doi.org/10.1177/2515245917745058.

Sadnicka, A., Strudwick, A.-M., Grogan, J. P., Manohar, S., & Nielsen, G. (2024). Going 'meta': a systematic review of metacognition and functional neurological disorder. *Brain Communications*, *7*(1), http://dx.doi.org/10.1093/braincomms/fcaf014.

Sánchez-Fuenzalida, N., van Gaal, S., Fleming, S. M., Haaf, J. M., & Fahrenfort, J. J. (2025). Confidence reports during perceptual decision making dissociate from changes in subjective experience. *Communications Psychology*, *3*(1), http://dx.doi.org/10.1038/s44271-025-00257-y.

Schwarz, W. (2022). *Random Walk and Diffusion Models: An Introduction for Life and Behavioral Scientists*. Springer International Publishing, http://dx.doi.org/10.1007/978-3-031-12100-5.

Shinn, M., Lam, N. H., & Murray, J. D. (2020). A flexible framework for simulating and fitting generalized drift-diffusion models. *ELife*, *9*, http://dx.doi.org/10.7554/elife.56938.

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-n design. *Psychonomic Bulletin & Review*, *25*(6), 2083–2101. http://dx.doi.org/10.3758/s13423-018-1451-8.

Softky, W. R., & Koch, C. (1993). The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *Journal of Neuroscience*, *13*(1), 334–350. http://dx.doi.org/10.1523/JNEUROSCI.13-01-00334.1993.

Stockart, F., Msheik, R., Robin, A., Jurkovičová, L., Goueytes, D., Rouy, M., Mareček, R., Hoffmann, D., Mudrik, L., Roman, R., et al. (2024). Cortical evidence accumulation for perceptual experience occurs irrespective of reports. BioRxiv, Retrieved from https://doi.org/10.1101/2024.03.20.585198.

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, *25*(3), 251–260. http://dx.doi.org/10.1007/bf02289729.

Talluri, B. C., Urai, A. E., Tsetsos, K., Usher, M., & Donner, T. H. (2018). Confirmation bias through selective overweighting of choice-consistent evidence. *Current Biology*, *28*(19), 3128–3135.e8. http://dx.doi.org/10.1016/j.cub.2018.07.052.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592. http://dx.doi.org/10.1037/0033-295x.108.3.550.

Von Luxburg, U., & Franz, V. H. (2009). A geometric approach to confidence sets for ratios: Fieller's theorem, generalizations and bootstrap. *Statistica Sinica*, 1095–1117.

Voss, A., Lerche, V., Mertens, U., & Voss, J. (2019). Sequential sampling models with variable boundaries and non-normal noise: A comparison of six models. *Psychonomic Bulletin & Review*, *26*(3), 813–832. http://dx.doi.org/10.3758/s13423-018-1560-4.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

Wald, A., & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, 326–339.

West, R. K., A-Izzeddin, E. J., Sewell, D. K., & Harrison, W. J. (2025). Priors for natural image statistics inform confidence in perceptual decisions. *Consciousness and Cognition*, *128*, Article 103818. http://dx.doi.org/10.1016/j.concog.2025.103818.

West, R. K., Harrison, W. J., Matthews, N., Mattingley, J. B., & Sewell, D. K. (2023). Modality independent or modality specific? Common computations underlie confidence judgements in visual and auditory decisions. *PLoS Computational Biology*, *19*(7), Article e1011245. http://dx.doi.org/10.1371/journal.pcbi.1011245.

Whiteley, L., & Sahani, M. (2008). Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes. *Journal of Vision*, *8*(3), 2. http://dx.doi.org/10.1167/8.3.2.

Wieschen, E. M., Voss, A., & Radev, S. (2020). Jumping to conclusion? A Lévy flight model of decision making. *The Quantitative Methods for Psychology*, *16*(2), 120–132. http://dx.doi.org/10.20982/tqmp.16.2.p120.

Yang, T., & Shadlen, M. N. (2007). Probabilistic reasoning by neurons. *Nature*, *447*(7148), 1075–1080. http://dx.doi.org/10.1038/nature05852.

Zhang, H., & Maloney, L. T. (2012). Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience*, *6*, 1. http://dx.doi.org/10.3389/fnins.2012.00001.

Zhang, H., Ren, X., & Maloney, L. T. (2020). The bounded rationality of probability distortion. *Proceedings of the National Academy of Sciences*, *117*(36), 22024–22034. http://dx.doi.org/10.1073/pnas.1922401117.

Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience*, *6*, 79. http://dx.doi.org/10.3389/fnint.2012.00079.

Zylberberg, A., Fetsch, C. R., & Shadlen, M. N. (2016). The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision. *ELife*, *5*, Article e17688. http://dx.doi.org/10.7554/eLife.17688.