Confidence in masked orientation judgments is informed by both evidence and visibility

Manuel Rausch^{1,2} · Sebastian Hellmann¹ · Michael Zehetleitner^{1,2}

Published online: 17 October 2017
© The Psychonomic Society, Inc. 2017

Abstract How do human observers determine their degree of belief that they are correct in a decision about a visual stimulus—that is, their confidence? According to prominent theories of confidence, the quality of stimulation should be positively related to confidence in correct decisions, and negatively to confidence in incorrect decisions. However, in a backward-masked orientation task with a varying stimulus onset asynchrony (SOA), we observed that confidence in incorrect decisions also increased with stimulus quality. Model fitting to our decision and confidence data revealed that the best explanation for the present data was the new weighted evidence-and-visibility model, according to which confidence is determined by evidence about the orientation as well as by the general visibility of the stimulus. Signal detection models, postdecisional accumulation models, two-channel models, and decision-time-based models were all unable to explain the pattern of confidence as a function of SOA and decision correctness. We suggest that the metacognitive system combines several cues related to the correctness of a decision about a visual stimulus in order to calculate decision confidence.

Keywords Metacognition · Confidence · Signal detection theory · Cognitive modeling · Masking · Perceptual decision making · Math modeling · Visual perception

Introduction

Humans are usually faced with the need to make decisions about stimuli on the basis of distorted or ambiguous sensory signals. To deal with that uncertainty, various strategies exist, such as looking the stimulus for a longer time, or asking other persons about their opinions. To know when these strategies are appropriate, observers need to have a sense for the probability that their decisions will be correct. The resulting subjective belief that a decision is correct is what we here refer to as *confidence* (Pouget, Drugowitsch, & Kepecs, 2016). To assess confidence experimentally, the standard procedure is to present observers in several trials with always one out of a set of two possible stimuli. Observers are asked to make a decision which of the two stimuli occurred and to indicate how confident they are about the accuracy of that decision (Baranski & Petrusic, 1994; Fleming & Dolan, 2012; Kepecs & Mainen, 2012). How do observers determine their degree of belief that a decision is correct? Although numerous different theories have been proposed in recent years (Fleming & Daw, 2017; Jang, Wallsten, & Huber, 2012; Kiani, Corthell, & Shadlen, 2014; Maniscalco & Lau, 2016; Moran, Teodorescu, & Usher, 2015; Paz, Insabato, Zylberberg, Deco, & Sigman, 2016; Pleskac & Busemeyer, 2010; Sanders, Hangya, & Kepecs, 2016), there is still no consensus as to the explanatory mechanism. In the present study, we suggest a novel alternative model for confidence in decisions about visual stimuli—namely, the *weighted-evidence-and-visibility model* (WEV model).

Signal detection theory (SDT) and related models

The most prominent class of models of confidence relies on signal detection theory (SDT; Green & Swets, 1966; Macmillan & Creelman, 2005; Wickens, 2002). SDT models assume the same mechanism underlying the decision, but they extend SDT in different ways to account for confidence.

✉ Manuel Rausch
manuel.rausch@ku.de

¹ Fakultät für Psychologie und Pädagogik, Fachgebiet Psychologie II, Katholische Universität Eichstätt-Ingolstadt, Ostenstraße 25, Eichstätt 85072, Germany

² Department Psychologie, Ludwig-Maximilians-Universität München, Munich, Germany

The mechanism proposed by SDT for making a decision is the following (Green & Swets, 1966; Macmillan & Creelman, 2005; Wickens, 2002): First, the stimulus generates sensory data within the visual system of the observer. Not all sensory data caused by the stimulus are relevant for the decision. Those aspects of the representation that are relevant for the task are transformed into sensory evidence, a continuous internal variable that differentiates between the two stimulus alternatives. Because there is noise in the system, the sensory evidence varies across presentations of the stimuli, which is why it is described as a random sample out of one of two probability distributions, one for each stimulus. If the observer is unable to differentiate between the stimuli, the two distributions created by the two stimuli are identical. The better the observer is able to differentiate between the two possible stimuli, the greater is the distance between the two distributions. It is assumed that observers, to select a response, compare their sample of evidence against a set of criteria. They respond “A” if the sample is greater than the criterion, and “B” otherwise.

How can SDT be extended to account for confidence, as well? The various proposals can be sorted into three categories (see Fig. 1):

- (i) Decision and confidence are based on identical samples of the sensory evidence (*SDT rating model*).
- (ii) Confidence is based on the same sample of sensory evidence as the decision, but the sample available for confidence is distorted or is overlaid with noise (*noisy SDT models*).
- (iii) Confidence is based on a second sample of the sensory evidence (*two-channel models*).

According to Proposal (i), the SDT rating model, decision options and levels of confidence are considered to form an

ordered set of responses, such as “I am sure it is A,” “I am guessing A,” “I am guessing B,” and “I am sure it is B.” Each adjacent pair of response options is delineated by one criterion. Both the decision and confidence are selected by the comparison of a single sample of evidence with the set of criteria. This means that observers respond, “I am guessing B” if the sample falls between that criterion separating “I am guessing B” from “I am guessing A” and that criterion separating “I am guessing B” from “I am sure it is B.”

Proposal (ii) is shared by several models that we here collectively refer to as *noisy SDT models*. These models assume that confidence is informed by the sample of evidence used to select the decision, but that the sample is distorted, incomplete, or overlaid with noise. The degree of confidence is then determined by comparing the distorted sample against a set of confidence criteria. The different theories categorized as noisy SDT models imply different mechanisms why the evidence used for confidence is less informative than the evidence for the decision: First, the sample could be overlaid by unsystematic noise (Hebart, Schriever, Donner, & Haynes, 2016; Sanders et al., 2016). Second, the evidence might be reduced (Maniscalco & Lau, 2016). Finally, confidence could be based exclusively on the evidence in favor of the selected option, but not on evidence against it (Zylberberg, Bartfeld, & Sigman, 2012).

Proposal (iii) is a tenet of models that we here refer to as *two-channel models*. These models were motivated by the observation that observers occasionally realize that they have made an error, which is not possible within the SDT rating model or the noisy SDT model (Fleming & Daw, 2017; Yeung & Summerfield, 2012). Two-channel models are characterized by the assumption that confidence is not based on the sample of evidence used to make the decision. Instead, it is based on a second sample of evidence generated in parallel to first. Although the term *two-channel model* seems to suggest that

Fig. 1 Three different extensions of the original signal detection model to account for confidence judgments.

the parallel samples are generated by two separate processes (Maniscalco & Lau, 2016), it is also possible that the same process generates first the sample for the decision and afterward the sample for confidence (Moran et al., 2015). Observers are then confident to the degree that the second sample confirms the decision. Although it is sometimes implied that the evidence used for confidence is stochastically independent from that used in the decision (Rausch & Zehetleitner, 2017), other models allow for a correlation between the two samples (Fleming & Daw, 2017; Jang et al., 2012).

The interaction between the quality of the stimulus and the correctness of the decision

How can SDT models of confidence be tested? A benchmark test may be found in the interaction between of the quality of stimulation and the correctness of the decision (Moran et al., 2015). Perceptual tasks can be made harder or easier by adjusting the physical features of the stimulus—for example, contrast, luminance, or presentation time. Those experimental manipulations intended to facilitate or complicate the task are what we collectively refer to here as the *quality of stimulation*. SDT and many models that assume that confidence is based on evidence predict the same qualitative pattern (Kellen & Klauer, 2015; Kepecs, Uchida, Zariwala, & Mainen, 2008; Sanders et al., 2016; Urai, Braun, & Donner, 2017): When the decision is correct, confidence should be positively associated with the stimulus quality. In contrast, when the decision is incorrect, the correlation between confidence and the stimulus quality should be negative. The reason for the interaction can be seen in Fig. 2: To inspire confidence, the sample of evidence must be more extreme than the criteria separating

decisions without confidence from decisions with confidence. When the stimulus quality is low, and thus the two stimuli are very hard to distinguish, the two distributions of evidence overlap almost entirely (left panel). When the stimulus quality is better, the two distributions are shifted away from each other (right panel). Therefore, greater portions of the distributions extend beyond the criteria for confidence in the correct decision, implying that confidence in the correctness of a decision becomes more and more likely. However, when the distance between the two distributions is large, the portions of the distributions that exceed the confidence criteria for incorrect decisions, which are located at the other sides of the distributions, become smaller and smaller (highlighted in black). As a consequence, when the quality of stimulation increases, confidence in incorrect decisions will become less likely.

Is there empirical support for such an interaction pattern? Previous studies have concurrently observed the predicted positive correlation between stimulus quality and confidence in *correct decisions* across a variety of tasks (Kiani et al., 2014; Moran et al., 2015; Rausch & Zehetleitner, 2016; Sanders et al., 2016). However, these studies were inconsistent regarding the predicted negative correlation between stimulus quality and confidence in *incorrect decisions*: The correlation was negative, as predicted by SDT, in an auditory discrimination task and a general-knowledge task (Sanders et al., 2016), as well as in a discrimination task concerning the relative amounts of white and black areas in a visual texture (Moran et al., 2015). However, in contrast to the SDT prediction, coherence of motion was positively, not negatively, associated with confidence in incorrect trials in a random-dot motion discrimination task (Kiani et al., 2014). Furthermore, in a low-contrast orientation discrimination task, the average correlation

Fig. 2 Distributions of evidence when the quality of the stimulus is low (left panel) and high (right panel). The areas in black highlight the probabilities of an incorrect decision while being confident.

between stimulus contrast and confidence in incorrect responses was close to zero with a narrow confidence interval (Rausch & Zehetleitner, 2016). Overall, the observations of both positive and negative correlations between stimulus quality and confidence in incorrect decisions pose a challenge to models of confidence derived from SDT.

The weighted evidence and visibility model

The present study proposes a new model of confidence in visual decisions, the weighted-evidence-and-visibility model (WEV). The core idea is that confidence judgments are influenced by two internal variables: the evidence as well as the visibility of the stimulus (see Fig. 3 and the Appendix for details).

The reason why visibility is informative for confidence is because stimuli always comprise several features: for instance, size, form, color, duration, and so forth. In many tasks, each stimulus falls into one of two categories, and observers need to decide which category the current stimulus belongs to. Only one or two features of the stimulus usually determine the category of the stimulus. Nevertheless, the visual system does not only represent the task-relevant, but also the task-irrelevant, features (Marshall & Bays, 2013; Xu, 2010). The strengths of the representations of task-relevant and task-irrelevant

features vary from trial to trial and may to some degree be independent of each other, as a consequence of the parallel processing of features in the visual system (Kyllingsbæk & Bundesen, 2007). To make a decision about the category of the stimulus, the representation of task-relevant features is transformed into evidence about the two stimulus categories. The representation of task-irrelevant features is not informative about the stimulus category, and therefore cannot contribute to the evidence. However, the representation of task-irrelevant features is not entirely useless, because the strength of this representation is informative about the quality of stimulation. This is particularly true for experiments in which the category of the stimulus and the quality of stimulation are both experimentally varied across trials. When many features of the stimulus are highly visible, it is reasonable to assume that the representation of the task-relevant feature will be accurate, and thus a high degree of confidence would be appropriate. Likewise, when the task-irrelevant features of the stimulus cannot be perceived, it is likely that the representation of the task-relevant feature is also poor, and thus the evidence could be misleading.

Therefore, the WEV model assumes that confidence is based not only on evidence. The second source of input is an estimate of the physical quality of the stimulus, which we refer to here as the *visibility* of the stimulus. For computational simplicity, the WEV model assumes that evidence and visibility both depend on the quality of stimulation but are stochastically independent when the stimulus quality is held constant. To determine the degree of confidence, the evidence and visibility of the stimulus are weighted and combined into one decision variable. The weights between evidence and visibility are expected to depend on the characteristics of the stimulation and the task: Some stimulus materials may allow observers to estimate the quality of stimulation with some precision. In this case, a strong weight on visibility would be expected. Since the visibility of the stimulus is determined by the quality of stimulation, but not by the category of the stimulus, the consequence is a positive correlation between stimulus quality and confidence in incorrect decisions. In contrast, other stimulus materials may leave observers without any cues to estimate the quality of stimulation, resulting in a strong weight on evidence. A strong weight on evidence implies a negative correlation between stimulus quality and confidence during incorrect decisions. Overall, the weighting of evidence and visibility allows the WEV model to be consistent with both positive and negative correlation patterns.

Fig. 3 The weighted evidence-and-visibility model. The stimulus varies in two aspects: the stimulus category (symbolized here as “A” and “B”) and a manipulation of stimulus quality (symbolized by the noise dots). The stimulus is assumed to generate two internal variables: the evidence, a continuous variable that differentiates between the categories “A” and “B,” and the visibility, a continuous variable that differentiates between “no visibility of the stimulus” and “full visibility.” The evidence is used to make a decision about the category of the stimulus. Confidence about the correctness of the decision, however, is based on both evidence and visibility.

Confidence and decision time

An alternative explanation for positive correlations between stimulus quality and confidence in incorrect decisions is provided by two sequential-sampling models of decision making: the diffusion model with internal deadlines (Ratcliff, 1978)

and the bounded-accumulation model (Kiani et al., 2014). These models share the assumption that confidence is informed by the elapsed time before a decision is reached. According to the internal deadlines model, observers set themselves variable deadlines for making the decision: When the decision is made before any of the deadlines has passed, the observer feels maximally confident that the decision is correct. The more deadlines are missed, the less confident is the observer that the decision is correct (Ratcliff, 1978). Can internal deadlines explain positive correlations between stimulus quality and confidence in incorrect decisions? Because the model assumes that confidence is informed only by decision time, when increasing the stimulus quality speeds up the decision time for incorrect decisions, increasing the stimulus quality should also make observers more confident in incorrect decisions. In contrast, when increasing the stimulus quality slows down the decision time for incorrect decisions, increasing the stimulus quality should decrease confidence in incorrect decisions. As a consequence, the internal-deadlines model can be tested by comparing the correlation between stimulus quality and confidence with the correlation between reaction time and confidence. If the model was correct, the correlations should be of different signs.

The bounded-accumulation model implies a more complex relationship between decision times and confidence. There, it is argued that sensory evidence is accumulated within two processes. The first process to hit a decision boundary determines which of the two stimulus categories will be selected. Over time, observers learn to associate decision times and states of the losing accumulator with the probability of being correct. When observers make a confidence judgment, they compare the decision time and the state of the losing accumulator with the distributions of decision times and accumulator states of correct and incorrect decisions they have learned over time. When the current decision time and accumulator state are more likely to stem from the distributions known from correct trials than from those of incorrect trials, observers are confident that the decision is correct (Kiani et al., 2014; van den Berg, Anandalingam, et al., 2016; van den Berg, Zylberberg, Kiani, Shadlen, & Wolpert, 2016).

Can the bounded-accumulation model explain positive correlations between stimulus quality and confidence in incorrect decisions? The state of the losing accumulator alone cannot account for these positive correlations because, as it accumulates evidence, it implies a negative correlation between stimulus quality and confidence in incorrect decisions, just as SDT does. However, the positive correlations between stimulus quality and confidence in incorrect decisions can be explained by observers learning the distribution of decision times. There are two possibilities. First, observers might learn that correct decisions are faster than incorrect decisions. In this case, if increasing the stimulus quality speeds up decision times for incorrect decisions, confidence in these incorrect decisions

should be increased, as well. Second, observers could learn that correct decisions are slower than incorrect decisions. In this case, confidence in incorrect decisions should increase with stimulus quality only if increasing the stimulus quality slows down the decision times for incorrect decisions.

To summarize, models of confidence based on decision times are in principle consistent with positive correlations between stimulus quality and confidence in incorrect decisions. However, positive correlations between stimulus quality and confidence imply specific patterns of decision times, which can be tested.

Rationale of the present study

The present study was designed to test whether the WEV model provides a better account of confidence in visual decisions than previously established models of confidence. For this purpose, we presented participants with horizontal and vertical gratings. In Experiment 1, the participants reported both the orientation of the stimuli and their confidence in being correct with just one single response. In Experiment 2, confidence judgments were made after the orientation judgment. To manipulate the quality of stimulation, we varied the stimulus onset asynchrony (SOA) between the grating and a backward mask. The backward mask was intended to interfere with both the representation of the task-relevant stimulus feature—that is, the orientation—and the representation of task-irrelevant features, for example, the shape.

In both experiments, the associations between SOA and confidence and between SOA and response times were assessed separately for correct and incorrect decisions. Moreover, we formally assessed the goodness of fit of seven models fitted separately to the orientation judgments and confidence reports of each single participant. The seven models included the WEV model, the SDT model, the noisy SDT model, the postdecisional accumulation model, and the two-channel model. In addition, two variants of the WEV model and the SDT model were used: one model in which the variability of evidence increased as a function of stimulus quality, and one in which the variability remained constant. These two models with SOA-dependent variance of the evidence were included because pilot studies had indicated that allowing variance to increase with SOA improved the model fit. Although previous studies had determined the parameters of the postdecisional accumulation models based on judgments, confidence, and response times (Moran et al., 2015; Pleskac & Bussemeyer, 2010), in the present study we did not model response times, because the WEV and SDT models do not make predictions about response times, and model comparisons need to be made on the same data.

We hypothesized that if the WEV model provides a better account of confidence in masked orientation decisions than any of the other five competing models, the two variants of

the WEV model should result in better goodness-of-fit measures than any of the competing models. If confidence was informed only by evidence, as is argued by SDT and related models, the correlation between SOA and confidence in incorrect trials should be negative. A positive correlation of SOA and confidence in incorrect trials would be consistent with the WEV model. Finally, if the correlations between SOA and confidence in incorrect trials can be explained by confidence being informed by decision times, the correlations between SOA and confidence should be consistent to the correlation between SOA and response time.

Experiment 1

The full experimental program, the analysis code, and the full data are openly available at the Open Science Framework website (<https://osf.io/ty4h8>), to facilitate reproduction of the present study and its results (Ince, Hatton, & Graham-Cumming, 2012; Morin et al., 2012). In addition, the hypotheses and analyses, including the participant exclusion criteria, were recorded at the same website (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). At this point in time we had already collected the data, but they had not yet been analyzed.

Material and methods

Participants A total of 20 participants (three male, 17 female) took part in the experiment. The age of the participants ranged between 18 and 40 years ($M = 23.3$). All participants reported normal or corrected-to-normal vision, no history of neuropsychological or psychiatric disorders, and not being on psychoactive medications. All participants gave written informed consent and received either course credits or €8 per hour for participation.

Apparatus and stimuli The experiment was performed in a darkened room. The stimuli were presented on a Display++ LCD monitor (Cambridge Research Systems, UK) with a screen diagonal of 81.3 cm, set at a resolution of $1,920 \times 1,080$ pixels and a refresh rate of 120 Hz. The viewing distance, not enforced by restraints, was approximately 60 cm. The experiment was conducted using PsychoPy version 1.83.04 (Peirce, 2007, 2008) on a Fujitsu ESPRIMO P756/E90+ desktop computer with Windows 8.1. The target stimulus was a square (size $3^\circ \times 3^\circ$), textured with a sinusoidal grating with one cycle per degree of visual angle (maximal luminance: 64 cd/m^2 ; minimal luminance: 21 cd/m^2). The mask consisted of a square ($4^\circ \times 4^\circ$) with a black- (0 cd/m^2) and-white (88 cd/m^2) checkered pattern consisting of five columns and rows.

All stimuli were presented at fixation against a gray (44 cd/m^2) background. The orientation of the grating varied randomly between horizontal or vertical. The participants simultaneously reported the orientation of the target and their confidence in being correct using a Cyborg V1 joystick (Cyborg Gaming, UK). Confidence was recorded using a continuous scale because continuous scales provide a maximum amount of information per single trial (Rausch & Zehetleitner, 2014).

Trial structure The time course of one trial is shown in Fig. 4. Each trial began with the presentation of a fixation cross for 1,000 ms. Then the target stimulus was shown for a short period of time until it was replaced by the mask. There were five possible SOAs, or time periods between target and mask onset: 8.3, 16.7, 33.3, 66.7, and 133.3 ms. The mask was always presented for 500 ms. When it disappeared, two visual analog scales were displayed on the screen. The upper scale represented the response that the orientation of the target was vertical, and the bottom one represented the response that the orientation of the target was horizontal. Observers reported the orientation by moving the joystick forward or back, which moved an index from the lower to the upper scale or vice versa. The ends of both scales were labeled *unsure* and *sure*. Observers reported their confidence by moving the joystick toward the left or the right, thus moving the index on the scales. To avoid any bias by the initial position of the index, the index appeared only when observers had started moving the joystick. The orientation response and degree of confidence were recorded when observers pulled the trigger of the joystick. Finally, if the orientation response was wrong, the trial ended by presentation of the word *error* for 1,000 ms.

Design and procedure Participants were instructed to report the orientation of the grating and their confidence in being correct on the orientation judgment as accurately as possible without time pressure. The experiment consisted of one training block and nine experimental blocks of 60 trials each. Each SOA featured 12 times in each block, in random order. The orientation of the target stimulus varied randomly across trials. After each block, the percentage of errors was displayed in order to provide participants with feedback about their accuracy.

Analysis

All analyses were conducted using the free software R (R Development Core Team, 2014).

Fig. 4 Trial structure of Experiment 1.

Correlation analysis To assess the relationships of confidence with both SOA and response times, we calculated the correlations between confidence and the other variables separately for each participant and for correct and incorrect decisions. Correlations were measured by Goodman and Kruskal's gamma coefficient (Γ), because Γ does not make any scaling assumptions beyond the ordinal level and can attain its maximum value irrespective of ties (Nelson, 1984).

Modeling analysis We fitted seven different models derived from SDT to the joint distributions of orientation responses and confidence data:

- (i) the SDT rating model
- (ii) the noisy SDT model
- (iii) the SDT model with SOA-dependent variance of evidence
- (iv) the WEV model
- (v) the WEV model with SOA-dependent variance of evidence
- (vi) the two-channel model
- (vii) the postdecisional accumulation model

In all seven models, we assumed that each presentation of the stimulus created a sample of evidence x drawn from a Gaussian distribution. The location of the distribution was determined by the stimulus category $S \in \{0, 1\}$, as well as the sensitivity parameter d specific to each SOA. Thus, each model involved five different sensitivity parameters d_1 – d_5 , one for each SOA. The mean of the distribution was calculated as $(S - \frac{1}{2}) \times d_i$: Thus, when $S = 1$, the distribution is centered at $d/2$, and when $S = 0$, the center is at $-d/2$. In models (iii) and (v), the variance s^2 of the distributions of x increased as a function of d : $s^2 = 1 + k \times d_i^2$, where k is a free parameter quantifying the slope of the increase of variance with d .

These two models were included in the analysis because an analysis of pilot experiments revealed that allowing the variance of evidence to increase as a function SOA improved the fits of the SDT and WEV model. In all other models, the variance s^2 was set to 1. In all models, participants' responses R were assumed to be 0 when x was lower than the primary task criterion θ , and 1 otherwise.

The degree of confidence was determined by comparing the decision variable y with a set of confidence criteria c . Each confidence criterion delineated between two adjacent confidence categories—for example, participants would select Confidence Category 2 if y fell between c_1 (which separated Categories 1 and 2) and c_2 (which separated Categories 2 and 3). To be consistent with the standard SDT rating model (Green & Swets, 1966), two separate sets of confidence criteria were assumed, one for each response option. The different models were characterized by the way that y was determined:

- According to the SDT rating model and the SDT model with an SOA-dependent variance of evidence, y was identical to x .
- According to the noisy SDT model, y was sampled from a Gaussian distribution, with a mean of x and with the standard deviation σ , which is an additional free parameter.
- According to the WEV model and the WEV model with SOA-dependent variance of evidence, y was again sampled from a Gaussian distribution with the standard deviation σ . The mean of the distribution from which y is drawn was calculated as $(1 - w) \times x + w \times (2R - 1) \times (d_i - \text{mean}(d))$. The parameter w described the degree to which participants relied on evidence or on visibility when they reported their degree of confidence. When $w = 0$, then the formula reduced to $y = x$; meaning the model was then identical to the noisy SDT model. The closer w was to 1, the more y depended on the term $(2R - 1) \times (d_i - \text{mean}(d))$.

The term $d_i - \text{mean}(d)$ ensured that y depended on the quality of stimulation, as indexed by d_i , independent of x . The term $2R - 1$ ensured that highly visible stimuli tended to shift y in such a way that high-confidence responses were more likely, and likewise, low-visibility stimuli shifted y in such a way that the probability of low-confidence responses increased.

- According to the two-channel model, y was stochastically independently from x . The mean of the distribution of y was given by $a \times d_i$, where a is a free parameter that expresses the fraction of signal available to the second channel as compared to the signal available to the first channel. The variance of y was set at 1.
- According to the postdecisional accumulation model, y was again sampled from a Gaussian distribution. The mean of the distribution is given by $x + (2S - 1) \times d_i \times b$, where b indicated the amount of postdecisional accumulation, and the term $2S - 1$ ensured that postdecisional accumulation tended to decrease y when $S = 0$, and to increase y when $S = 1$. The variance of the distribution of y was b^2 .

Model fitting was performed separately for each single participant. The fitting procedure involved the following computational steps. First, the continuous confidence ratings were discretized by dividing the continuous scale into five equal partitions. Second, the frequency of each confidence category was calculated for both orientations of the stimulus and the orientation response. Third, for each model, the set of parameters was determined that minimized the negative log-likelihood of the data (Dorfman & Alf, 1969). The formulae for calculating the probability of an orientation response in conjunction with a specific degree of confidence given the stimulus and the set of parameters are described in the [Appendix](#). Minimization of the negative log-likelihood was performed using a general SIMPLEX minimization routine (Nelder & Mead, 1965).

The relative goodness of fit of the seven candidate models was assessed using the Bayes information criterion (Schwarz, 1978) and the AIC_c (Burnham & Anderson, 2002), a variant of the Akaike information criterion (Akaike, 1974).

Group-level statistics All statistical analysis at the group level was based on Bayesian statistics (Dienes, 2011; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wetzels et al., 2011) using the R library BayesFactor (Morey & Rouder, 2014). To test whether the mean Γ correlation coefficients were different from zero and to compare the BICs and AIC_c s between two models, we used the Bayesian equivalents of t tests. Recommended as a standard prior in psychology, a Cauchy distribution with a scale parameter of 1 was assumed as the prior distribution for standardized effect sizes (Rouder et al., 2009). In addition, Markov chain Monte Carlo resampling was used to determine the 95% credible intervals of the posterior distribution of the mean Γ as well as the mean differences for BIC and AIC_c . In addition, the Bayesian equivalent of an analysis of variance was performed with confidence as the dependent variable and with SOA and trial correctness as factors, testing all models that can be created by removing or leaving in a main effect or interaction term from the full model. As priors, we used a default g prior with a scale parameter of $\sqrt{2}/2$ to maintain consistency with the prior of the Bayesian t tests (see Rouder & Morey, 2012, for the model details).

Results

Two participants were excluded from the analysis because their error rates were not below chance level, $BF_{10} \leq .20$. For the remaining 18 participants, the error rates were at chance at the SOA of 8.3 ms ($M = 50.3\%$, $SD = 6.5$) and dropped to $M = 2.6\%$ ($SD = 6.7$) at the maximum SOA of 133.3 ms (see Fig. 5, left panel). Confidence judgments averaged 11.5% ($SD = 11.7$) of the width of the visual analog scale at an SOA of 8.3 ms and increased to a mean of 90.2% ($SD = 14.0$) at an SOA of 133.3 ms.

Correlation between SOA and confidence As can be seen from the central panel of Fig. 5, confidence increased as a function of SOA in correct as well as in incorrect trials, although to a lesser extent in the latter case. The Bayesian equiv-

Fig. 5 Mean error rates (left panel), mean confidence (central panel), and mean response times (right panel), as a function of stimulus onset asynchrony (SOA; x-axes). Error bars = 1 SEM.

alent of an analysis of variance revealed effects of SOA, $BF_{10} = 2.1 \cdot 10^{45}$, and trial correctness on confidence, $BF_{10} = 2.3 \cdot 10^3$, as well as an interaction between the two, $BF_{10} = 31.4$. The mean correlation coefficient between confidence and SOA was large for correct trials, $M = .80$, $SD = .08$, and medium for incorrect trials, $M = .38$, $SD = .19$. A Bayesian analysis indicated that the mean correlation coefficients were different from zero for correct trials, posterior of the mean 95% CI = [.76, .84], $BF_{10} = 3.57 \cdot 10^{15}$, as well as for incorrect trials, posterior of the mean 95% CI = [.27, .46], $BF_{10} = 8.77 \cdot 10^4$.

Correlation between SOA and response times As is shown in the right panel of Fig. 5, correct responses were faster than incorrect responses, posterior of the mean RT difference 95% CI = [54.9, 262.2], $BF_{10} = 11.3$. Correct responses also appeared to become faster with increasing SOA. However, the mean correlation was only weakly negative, $M = -.11$, $SD = .26$, and there was insufficient evidence to ascertain whether the mean correlation was different from 0, posterior of the mean 95% CI = [−.22, .02], $BF_{10} = .74$. Concerning incorrect

responses, the right panel of Fig. 5 suggests that response times remained constant for shorter SOAs between 8.3 and 33.0 ms, and sharply increased for the longer SOAs. However, the mean correlation was negligibly small, $M = .05$, $SD = .14$. The reason is that the number of errors at longer SOAs was low, which is why errors at shorter SOAs contributed more strongly to the correlation coefficients. Consequently, there was insufficient evidence to ascertain whether the mean correlation was zero, posterior of the mean 95% CI = [−.02, .11], $BF_{10} = .49$.

Modeling results As is suggested by the upper left and upper central panels of Fig. 6, the only two models that could reproduce the increase of confidence with SOA in incorrect trials were the WEV model and the WEV model with SOA-dependent variance of the evidence. All the other models predicted a decrease of confidence in incorrect trials with SOA, but no such a decrease was observed.

Figure 7 shows that the two flavors of the WEV model could reproduce the correlation between SOA and confidence

Fig. 6 Mean confidence predicted by the seven models (in separate panels) as a function of stimulus onset asynchrony (x-axis) and the correctness of the orientation response in Experiment 1. Solid lines

indicate predictions for correct trials, and dashed lines predictions for incorrect trials. Circles indicate the observed confidence in correct trials, and triangles that in incorrect trials. Error bars indicate 1 SEM.

Fig. 7 Observed gamma correlation coefficients between SOA and confidence as a function of the predicted gamma correlation coefficients for the different models in Experiment 1, in separate panels. Each symbol

represents the data from one participant. Error bars indicate the standard errors of the correlation coefficients.

at the level of each single participant. The other models systematically failed to account for confidence in incorrect decisions.

The two WEV models also fitted the data best in terms of BIC and AIC_c : The best model overall was the WEV model with SOA-dependent variance of the evidence, BIC: $M = 1,386$, AIC_c : $M = 1,314$, followed by the WEV model with constant variance, BIC: $M = 1,388$, AIC_c : $M = 1,320$. The third-best model was the SDT model with SOA-dependent variance of evidence, but its fit to the data was worse according to both metrics (BIC: $M = 1,425$; AIC_c : $M = 1,362$). All the

other models performed even worse (all BICs: $M \geq 1,506$; all AIC_c s: $M \geq 1,447$).

A series of Bayes factors was used to compare the AIC_c and BIC of the WEV model with SOA-dependent variance of evidence against those of all six of the other models. The Bayes factors revealed some evidence against a difference in model fits between the WEV model with SOA-dependent variance and the WEV model with constant variance:

- BIC: posterior of the mean ΔBIC : 95% CI = $[-14.1, 11.4]$, $BF_{10} = 0.18$

- AIC_c : posterior of the mean ΔAIC_c : 95% CI = [– 18.1, 7.6], $BF_{10} = 0.25$.

However, the Bayes factors indicated strongly that the WEV model with SOA-dependent variance fitted the data better than each of the other five models:

- SDT model with SOA-dependent variance of evidence:
 - BIC: posterior of the mean ΔBIC : 95% CI = [– 55.1, – 18.9], $BF_{10} = 109.2$
 - AIC_c : posterior of the mean ΔAIC_c : 95% CI = [– 63.6, – 27.1], $BF_{10} = 695.8$
- SDT rating model:
 - BIC: posterior of the mean ΔBIC : 95% CI = [– 162.9, – 67.6], $BF_{10} = 546.9$
 - AIC_c : posterior of the mean ΔAIC_c : 95% CI = [– 175.3, – 79.3], $BF_{10} = 1.5 \cdot 10^3$
- postdecisional accumulation model:
 - BIC: posterior of the mean ΔBIC : 95% CI = [– 167.9, – 71.8], $BF_{10} = 801.4$
 - AIC_c : posterior of the mean ΔAIC_c : 95% CI = [– 176.8, – 79.9], $BF_{10} = 1.6 \cdot 10^3$
- noisy SDT model:
 - BIC: posterior of the mean ΔBIC : 95% CI = [– 168.8, – 73.0], $BF_{10} = 897.6$
 - AIC_c : posterior of the mean ΔAIC_c : 95% CI = [– 177.5, – 81.2], $BF_{10} = 1.8 \cdot 10^3$
- two-channel model:
 - BIC: posterior of the mean ΔBIC : 95% CI = [– 177.4, – 80.3], $BF_{10} = 1.5 \cdot 10^3$
 - AIC_c : posterior of the mean ΔAIC_c : 95% CI = [– 184.7, – 88.8], $BF_{10} = 3.0 \cdot 10^3$.

In view of these results, we performed an additional simulation to assess whether the present results can be explained by a failure of model identification. For this purpose, we simulated data based on the SDT rating model and the SDT model with SOA-dependent variance of evidence (i.e., the two models that received the most empirical support without assuming an influence of visibility weighting). We first created 1,000 bootstrap samples from the parameter sets obtained by fitting the empirical data. Then, for each sampled parameter set, we randomly created a data set with the same number of trials as the empirical data. Using each simulated data set, we

fitted both the WEV model with SOA-dependent variance and the correct generative model—that is, the SDT model or the SDT model with SOA-dependent variance, respectively. Comparing model fits between the WEV model with SOA-dependent variance and the SDT model based on data conforming to the SDT model revealed not a single BIC difference erroneously in favor of the WEV model with SOA-dependent variance, $f = 0.0\%$, and only a few misleading AIC_c differences, $f = 3.6\%$. Likewise, comparing the WEV model with SOA-dependent variance and the SDT model with SOA-dependent variance using data in line with the latter model suggested hardly any BIC differences incorrectly favoring the WEV model with SOA-dependent variance, $f = 0.3\%$, and a moderate number of misleading AIC_c differences, $f = 7.7\%$.

Discussion

The present experiment suggested that the WEV model with SOA-dependent variance of evidence provides a better account of confidence in masked orientation decisions than the SDT model, the noisy SDT model, the postdecisional accumulation model, and the two-channel model. Moreover, longer SOAs increased confidence in both correct and incorrect responses, an observation consistent with the WEV model but not in accordance with the SDT rating model and the many other confidence models. Concerning response times, there was insufficient evidence to determine whether SOA and response time were correlated in correct or incorrect trials.

Can the positive correlation between SOA and confidence in incorrect decisions be explained by the WEV model alone, or are models based on decision times able to provide an alternative explanation? Both the internal-deadline model and the bounded accumulation model could only account for the confidence data if there were a negative correlation between SOA and response times. The reason is that according to the internal-deadline model, when observers are more confident, this means that the decision process was faster, so fewer internal deadlines were missed. According to the bounded accumulation model, because correct responses were faster than incorrect responses, observers would associate fast responses with a greater probability of a correct response, meaning that confidence in incorrect decisions would increase with SOA only if response times decreased with SOA. Unfortunately, the data were inconclusive as to whether the correlation of RT and SOA was absent, although the credible interval indicated that if there were a negative correlation, its size would be close to zero. Nevertheless, since there was still the possibility of a correlation between SOA and response times, the deadline diffusion model and the bounded accumulation model could not be ruled out as explanations of the positive correlation between SOA and confidence in incorrect decisions in this experiment. There were several possibilities why the correlation between SOA and response times might

have remained undetected. First, it would be expected that longer SOAs would speed up correct responses. However, once again there was insufficient evidence for such a correlation. This might indicate that the assessment of response times is not very precise. An obstacle to assessing response times with precision was that observers had to make two decisions at the same time, one about the orientation of the stimulus and one about their degree of confidence. As a consequence, the response times might reflect not only the time needed to make a decision about the orientation, but also the time required to select a degree of confidence. The decision time related to confidence might have obscured the correlation between SOA and decision time related to the decision about the stimulus (Kiani et al., 2014). For this reason, we conducted a second experiment, instructing observers to report their degree of confidence only after the orientation response.

Is it possible that the simultaneous measurement of confidence and orientation responses interfered with the reported degree of confidence as well, not only with its timing? In line with this hypothesis, the other study to observe a positive correlation between confidence in incorrect decisions and stimulus quality had assessed confidence and task response with one single response, as well (Kiani et al., 2014). Those studies to find a negative correlation had all assessed confidence only after the task response (Moran et al., 2015; Sanders et al., 2016). These previous studies had offered two explanations why measuring confidence simultaneously with or subsequent to the decision might result in different correlations: One possibility is that assessing confidence only after the response might allow observers to accumulate additional sensory evidence (Kiani et al., 2014). Because sensory evidence implies negative correlations between stimulus quality and confidence in incorrect decisions, assessing confidence after the decision might miss a positive correlation between confidence and task accuracy at the time of the decision. A second possibility is that asking observers to report the orientation and their confidence at the same time might induce a decision strategy based no longer on evidence, but on heuristics (Aitchison, Bang, Bahrami, & Latham, 2015). Again, a

second experiment seemed necessary to investigate whether the results of Experiment 1 would generalize to an experiment in which observers reported their degree of confidence only after the orientation response.

Experiment 2

The materials, methods, analysis, preregistration, and availability of all materials of Experiment 2 were the same as for Experiment 1, except for the differences outlined below.

Material and methods

Participants A total of 39 participants (12 male, 27 female) took part in the experiment. The age of the participants ranged from 18 to 24 years ($M = 20.1$).

Trials structure A trial in Experiment 2 was the same as one in Experiment 1, except for the following difference (see Fig. 8). The mask was presented until participants gave a nonspeeded response by key press as to whether the target had been horizontal or vertical. Only after this response did participants report their confidence in being correct about the orientation. For this purpose, the question “How confident are you that your response was correct?” was displayed on screen. Participants reported their confidence using one visual analog scale and again a joystick, meaning that they selected a position along a continuous line between two endpoints by moving a cursor. The end points were labeled *unsure* and *sure*.

Results

Six participants were excluded from the analysis because their error rates were not below chance level, $BF_{10s} \leq .35$. For the 33 remaining participants, error rates were at chance level at the SOA of 8.3 ms ($M = 50.0\%$, $SD = 4.2$) and decreased to $M = 4.7\%$ ($SD = 5.5$) at the maximum SOA of 133.3 ms (see Fig. 9, left panel). Confidence averaged 15.3% ($SD = 18.3$) at the SOA

Fig. 8 Trial structure of Experiment 2.

Fig. 9 Mean error rates (left panel), mean confidence (central panel), and mean response times (right panel) as a function of stimulus onset asynchrony (SOA; x-axis). Error bars = 1 SEM.

of 8.3 ms and increased to a mean of 90.5% ($SD = 12.4$) at an SOA of 133.3 ms.

Correlation between SOA and confidence As can be seen in the central panel of Fig. 9, confidence increased as a sigmoid function of SOA in correct trials. For incorrect trials, confidence increased in the range between 8.3 and 66.7, but then appears to have decreased at the final SOA of 133 ms. The Bayesian equivalent of an analysis of variance on confidence indicated effects of SOA, $BF_{10} = 3.1 \cdot 10^{62}$, and correctness, $BF_{10} = 7.8 \cdot 10^{17}$, as well as an interaction between SOA and correctness, $BF_{10} = 5.4 \cdot 10^{16}$. The mean correlation coefficient between confidence and SOA was again large for correct trials, $M = .78$, $SD = .21$, and medium for incorrect trials, $M = .39$, $SD = .29$. A Bayesian analysis indicated that the mean correlation coefficients were different from zero for correct trials, posterior of the mean 95% CI = [.70, .85], $BF_{10} = 3.29 \cdot 10^{17}$, as well as for incorrect trials, posterior of the mean 95% CI = [.28, .49], $BF_{10} = 1.52 \cdot 10^6$. Given the apparent dip in confidence at the longest SOA, we decided post hoc to perform a test whether confidence decreased between the SOAs of 66.7 and 133.0. However, the Bayes factor was not conclusive: posterior of the mean decrease of confidence between the two SOAs 95% CI = [− 0.9, 25.4], $BF_{10} = 0.84$.

Correlation between SOA and response times The Bayes factor indicated that correct responses were faster than incorrect ones, posterior of the mean difference 95% CI = [69.4, 171.2], $BF_{10} = 1.0 \cdot 10^3$. As is shown in the right panel of Fig. 9, correct responses also became faster with increasing SOA. Although the mean correlation of SOA and response times in correct trials was $M = -.17$ ($SD = .20$), and thus similar in size to the one observed in Experiment 1, the Bayes factor in Experiment 2 indicated decisively that the mean correlation between SOA and response times in correct trials was different from zero, posterior of the mean 95% CI = [− .23, − .09], $BF_{10} = 635.5$. Just as in Experiment 1, response times to incorrect responses appeared to remain constant for shorter SOAs between 8.3 and

33.0, and to increase only for longer SOAs. Nevertheless, the mean correlation was negligibly small again, $M = .03$, $SD = .15$. The Bayesian analysis indicated some evidence that the mean correlation was zero: posterior of the mean 95% CI = [− .02, .08], $BF_{10} = .25$.

Modeling analysis As can be seen from the upper left and upper central panels of Fig. 10, the WEV model, in either the version with SOA-dependent variance of evidence or the version with constant variance, provided good qualitative accounts of the data of Experiment 2, except that they could not reproduce the confidence in incorrect trials at the maximum SOA of 133.3 ms.

The upper row of Fig. 11 again shows that the two flavors of the WEV model performed best in reproducing the correlation between SOA and confidence at the level of each single participant. Nevertheless, it can also be seen that most symbols are located to the right of the major diagonal, indicating that the two WEV models tended to overestimate the correlation between SOA and confidence. The SDT model with SOA-dependent variance was the only alternative model that occasionally reproduced a positive correlation between SOA and confidence in incorrect trials, whereas all the other models incorrectly predicted negative correlations between SOA and confidence in incorrect decisions.

The WEV models again were the two models that fitted the data best in terms of BIC and AIC_c : The best model overall was again the WEV model with SOA-dependent variance of evidence (BIC: $M = 1,406$; AIC_c : $M = 1,334$), followed by the WEV model with constant variance (BIC: $M = 1,410$, AIC_c : $M = 1,342$). The best model that did not involve weighting of evidence and visibility was the SDT model with SOA-dependent variance of evidence, although the fit was decisively worse according to both metrics (BIC: $M = 1,422$; AIC_c : $M = 1,359$). All the other models performed even worse (all BICs: $M \geq 1,529$; all AIC_c s: $M \geq 1,466$).

Just as in Experiment 1, the Bayes factor analysis revealed some evidence against a difference in model fits between the WEV model with SOA-dependent variance and the WEV

Fig. 10 Mean confidence predicted by the seven models (in separate panels), as a function of stimulus onset asynchrony (x-axis) and the accuracy of the orientation response in Experiment 2. Solid lines

indicate predictions for correct trials, and dashed lines those for incorrect trials. Circles indicate the observed confidence in correct trials, and triangles that in incorrect trials. Error bars indicate 1 SEM.

model with constant variance, although in Experiment 2 the result regarding the AIC_c was not conclusive:

- BIC: posterior of the mean ΔBIC : 95% CI = [− 13.6, 5.5], $BF_{10} = .20$,
- AIC_c : posterior of the mean ΔAIC_c : 95% CI = [− 17.5, 1.5], $BF_{10} = .55$.

Concerning the other models, the Bayes factors suggested that the WEV model with SOA-dependent variance fitted the data better than each of the five other models:

- SDT model with SOA-dependent variance:
 - BIC: posterior of the mean ΔBIC : 95% CI = [− 26.6, − 4.8], $BF_{10} = 6.3$,
 - AIC_c : posterior of the mean ΔAIC_c : 95% CI = [− 34.9, − 12.8], $BF_{10} = 287.2$;
- SDT rating model:

- BIC: posterior of the mean ΔBIC : 95% CI = [− 161.3, − 82.4], $BF_{10} = 6.4 \cdot 10^4$,
- AIC_c : posterior of the mean ΔAIC_c : 95% CI = [− 173.9, − 95.2], $BF_{10} = 3.7 \cdot 10^5$;

- postdecisional accumulation model:

- BIC: posterior of the mean ΔBIC : 95% CI = [− 160.7, − 79.6], $BF_{10} = 3.0 \cdot 10^4$,
- AIC_c : posterior of the mean ΔAIC_c : 95% CI = [− 168.9, − 87.6], $BF_{10} = 9.4 \cdot 10^4$;

- noisy SDT model:

- BIC: posterior of the mean ΔBIC : 95% CI = [− 168.2, − 88.6], $BF_{10} = 1.4 \cdot 10^5$,
- AIC_c : posterior of the mean ΔAIC_c : 95% CI = [− 176.3, − 96.7], $BF_{10} = 4.5 \cdot 10^5$;

- two-channel model:

Fig. 11 Observed gamma correlation coefficients between SOA and confidence as a function of the predicted gamma correlation coefficients for the different models in Experiment 2, in separate panels. Each symbol

represents the data from one participant. Error bars indicate the standard errors of the correlation coefficients.

- BIC: posterior of the mean ΔBIC : 95% CI = $[-168.9, -85.2]$, $\text{BF}_{10} = 5.2 \cdot 10^4$,
- AIC_c : posterior of the mean ΔAIC_c : 95% CI = $[-177.6, -93.8]$, $\text{BF}_{10} = 1.6 \cdot 10^5$.

Finally, we again performed a model identification analysis analogous to that in Experiment 1. Comparing model fits on data conforming to the SDT model between the SDT model and the WEV model with SOA-dependent variance again revealed not a single BIC comparison that incorrectly favored the WEV model with SOA-dependent variance, $f = 0.0\%$, and only few cases with a misleading AIC_c comparison, $f = 3.4\%$.

Similarly, a comparison between the WEV model with SOA-dependent variance and the SDT model with SOA-dependent variance using data in line with the latter model indicated a few cases with a BIC comparison erroneously in favor of the WEV model with SOA-dependent variance, $f = 3.4\%$, and a moderate number of instances with misleading AIC_c comparisons, $f = 6.8\%$.

Discussion

Experiment 2 suggested that the WEV model with SOA-dependent variance of evidence provides a better account of

confidence in masked orientation decisions than the SDT model with SOA-dependent variance of evidence, the SDT rating model, the noisy SDT model, the postdecisional accumulation model, and the two-channel model, even when confidence was recorded after orientation responses. Moreover, longer SOAs increased confidence in both correct and incorrect trials. Although the WEV model and the SOA-dependent SDT model could produce positive correlations between SOA and confidence in incorrect decisions, only the first one accounted reasonably well for confidence as a function of SOA. Concerning response times, SOAs speeded up response times in correct trials only, whereas in incorrect trials, there seemed to be no relation between response times and SOA.

Models based on response times cannot explain the positive correlation between SOA and confidence in incorrect trials that we observed in this experiment. The reason is that there was evidence that SOA and response time in incorrect trials were uncorrelated. If decision times had an effect on confidence, as is suggested by some models (Kiani et al., 2014; Ratcliff, 1978), the effect of decision times on confidence in incorrect trials should be approximately constant across SOAs, because decision times in incorrect decisions were also more-or-less constant across SOAs. Consequently, decision times cannot explain why confidence in incorrect decisions *increased* with SOA. It should be noted that the present study does not imply that there is *no* effect of decision time on confidence: What the present experiment does imply is that if there were an effect of decision time on confidence, this effect still would not explain the observed correlation between SOA and confidence in incorrect decisions in the present study.

Why was confidence in errors positively correlated with SOA in the present study, although previous studies had observed a negative correlation between confidence in errors and stimulus quality (Moran et al., 2015; Sanders et al., 2016)? Since virtually the same pattern of results occurred in Experiments 1 and 2, despite the fact that confidence and decisions were recorded simultaneously in Experiment 1, and decisions were recorded before confidence in Experiment 2, sequential or simultaneous assessment of decisions and confidence cannot explain the differences from previous studies (Kiani et al., 2014; Sanders et al., 2016). It might be speculated that some stimuli provide cues as to the quality of the stimulation, independent of the evidence, whereas evidence and cues as to the quality of the stimulation are identical in other stimuli. When no cues as to the quality of the stimulation are available, models based on evidence alone might be sufficient to describe confidence. Only when a percept is informative about the quality of stimulation independent of the evidence might the WEV model outperform the SDT model in explaining confidence. To identify those stimulus parameters that promote a high weighting of visibility, and thus a good fit to the WEV model, future studies will be necessary to measure confidence across a range of different tasks.

The apparent “dip” in confidence at the maximum SOA of 133.3 ms is an unresolved issue in Experiment 2: Although the WEV model predicts that confidence monotonically increases with SOA, in Experiment 2, confidence seemed to switch from increasing to decreasing at the maximum SOA. Because the statistical analysis of the dip was not conclusive, we could probably disregard the last data point as unreliable, due to the low number of errors at the maximum SOA, if we had not repeatedly observed a similar dip during pilot experiments. A potential explanation for the sudden decrease in confidence could be the detection of motor errors. Given that observers did detect that they had accidentally pressed the wrong button, they should have reported a minimal degree of confidence. However, confidence in incorrect decisions is challenging to measure at high SOAs because participants make hardly any errors at these SOAs. Future studies appear necessary to investigate whether the WEV model needs to be extended to encompass error detection in order to account for confidence in masked orientation decisions.

General discussion

The present study revealed that confidence in incorrect decisions in a masked orientation task is positively related to SOA. The effect was constant when the confidence and orientation responses were recorded either simultaneously (Exp. 1) or one after the other (Exp. 2). Moreover, Experiments 1 and 2 converged, in that the best fit to the data was obtained by the WEV model with SOA-dependent variance of evidence. In contrast to the WEV model, the SDT model, the SDT model with SOA-dependent variance of evidence, the noisy SDT model, the postdecisional accumulation model, and the two-channel model all failed to provide an adequate account of confidence in incorrect decisions as a function of SOA. Finally, the relation between confidence in incorrect decisions and SOA was also not satisfactorily explained by decision times.

Is confidence based on heuristic inference?

Theories according to which confidence is calculated from evidence are frequently contrasted with heuristic theories (Aitchison, Bang, Bahrami, & Latham, 2015; Barthelmé & Mamassian, 2010; Moran et al., 2015). Can the present results be explained in terms of heuristic inference? A heuristic is classically defined as judgmental operation used as a substitute for a more complex judgment (Tversky & Kahneman, 1975). A hallmark of heuristic inference is that heuristic decisions ignore part of the information. A large amount of research showed that human decision making is often subject to heuristic simplification (Gigerenzer & Gaissmaier, 2011). A common example of judgments thought to be too complex for

rational decision making and thus prone for heuristics are judgments about probabilities. Since confidence is often defined as the subjective probability of being correct (Sanders et al., 2016), heuristic inference may be expected to be used in confidence as well.

One candidate for a heuristic involved in the present study is the perceived duration of the target stimulus: This would mean that observers ignored the sensory evidence about the orientation, and instead based their confidence only on the time the stimulus was presented before it was replaced by the mask. However, the data is not consistent with that view: If confidence was exclusively based on perceived duration, confidence should be the same in correct and incorrect decisions when the SOA is held constant. However, observers were more confident in correct decisions than in incorrect ones at several SOAs, suggesting that although perceived duration could be one of the features that contributed to the visibility of the stimulus, it cannot be the only source of information to confidence.

The WEV model may constitute an example of heuristic inference or not depending on the specific concept of heuristic inference. In the context of theories about confidence, previous authors are not consistent in what calculations underlying confidence qualify as heuristic inference. First, in a task in which two features of the stimulus determined the difficulty of the decision conjointly, heuristics were argued to depend only on single features of the stimulus, in opposition to calculations of confidence in which the two features of the stimulus were combined (Barthelmé & Mamassian, 2010). Others stated that according to heuristic theories, confidence is not calculated from the evidence extracted during the decision process (Moran et al., 2015). Finally, contrasting the Bayesian with the heuristic approach, it was argued that all transformations of the sensory input quality as heuristic except when the probability of being correct is explicitly calculated (Aitchison et al., 2015). The WEV model shares some, but not all purported characteristics heuristic inference: Consistent with heuristic inference, the WEV model does not involve explicit calculations of the probability of being correct, and confidence is not exclusively based on the evidence. However, in contrast to what heuristic inference is often described, evidence does inform confidence in the WEV model. Crucially, the WEV model does not involve a simplification as sensory evidence is assumed to be combined with, not replaced by the visibility of the stimulus. Overall, irrespective of which is most suitable concept of heuristic inference in the context of decision confidence, it is necessary to stress that the WEV model combines purported features of heuristic and computational theories.

Confidence: Informed by many signals

Although a combination of sensory evidence and stimulus visibility appears sufficient to explain confidence in the

present study, there is converging evidence that confidence in visual decisions is determined by many cues.

A mechanism similar to one described in the present study relies on the variability of evidence provided by the stimulation: When observers discriminated the average color of an array of colored shapes, confidence was not only determined by the distance of the average color to the category boundary, but was also affected by the variability of color across the array (Boldt, de Gardelle, & Yeung, 2017). Moreover, in a global motion discrimination task, confidence depended on the consistency of motion signals, although discrimination performance was equated (Spence, Dux, & Arnold, 2015). A second mechanism may be based on monitoring the time required to reach a decision. Although we found no evidence for such a mechanism in the present study, in a previous study researchers had manipulated the time needed to reach the decision in a global motion discrimination task while equating task performance, and had observed that decision time directly informed confidence (Kiani et al., 2014). Finally, observers are able to occasionally detect their own errors (Yeung & Summerfield, 2012). On the basis of SDT or the WEV model, it would not be possible to change one's mind about which response is the correct one, which is why there must be an additional mechanism to explain error detection. Previous studies have argued that the accumulation of evidence for responses continues even after the decision is made, which could account for the detection of errors (van den Berg, Anandalingam, et al., 2016; Yeung & Summerfield, 2012). In Experiment 2 there was some indication that error detection might have played a role at the maximum SOA, although the data were not conclusive here.

Overall, it is tempting to speculate about a common mechanism underlying these effects: A metacognitive system may calculate confidence on the basis of as many cues to the probability of being correct as are available in a specific task. The visibility of the stimulus, variance of colors, variability of motion signals, elapsed decision time, and detected errors might all be cues in that system. Consistent with a shared metacognitive system, confidence and error detection share one electrophysiological marker, the Pe component (Boldt & Yeung, 2015). However, more research appears to be necessary to investigate whether a common mechanism underlies confidence in all these tasks.

Dynamics of decisions and confidence

Although the WEV model provides a reasonable account of decisions and confidence in a masked orientation task, it would be even more satisfying to have one model explain not only decision and confidence, but also decision times simultaneously. However, the WEV model was not designed to account for decision times. Several previous studies have attempted to provide such a unified account (Kiani et al., 2014; Moran et al.,

2015; Pleskac & Busemeyer, 2010), but these theories are not consistent with the present data: The bounded accumulation model cannot explain why the pattern of response times as a function of SOA was inconsistent with the pattern of confidence. Postdecisional accumulation models predict a negative correlation between SOA and confidence in incorrect trials, and thus are also not in accord with the present data.

Although this was not the focus of the present study, future studies might be able to provide an extension of the WEV model in the temporal domain. Though sequential sampling processes are natural candidates to account for the dynamics of evidence (Ratcliff, Smith, Brown, & McKoon, 2016), to our knowledge no study has investigated the temporal dynamics of cues to being correct other than the sensory evidence. Such a dynamical extension of the WEV model should be able to account for decisions, confidence, and decision times in masked orientation decisions at the same time.

Summary

The best explanation for confidence in masked orientation decisions is the WEV model, which argues that confidence is calculated from sensory evidence as well as the visibility of the stimulus. Signal detection models, postdecisional accumulation models, two-channel models, and decision-time-based models are all not able to explain the pattern of confidence in incorrect decisions as a function of SOA. We suggest that the metacognitive system calculates one's confidence by combining multiple cues about being correct in visual decisions.

Author note We are grateful to Rani Moran and Marius Usher for making us aware of the problem of explaining positive correlations between SOA and confidence in incorrect trials. In addition, we thank Laura Fumy, Theresa Güthoff, Sandra Miller, and Lisa Tömm for help with the data collection.

Appendix

The WEV model with SOA-dependent variance of evidence

The WEV model with SOA-dependent variance of evidence is described here first, because it represents a general case that incorporates the SDT model, the SDT model with SOA-dependent variance, the noisy SDT model, and the WEV model with constant variance as special cases. The differences from the other models are explained below.

According to the WEV model with SOA-dependent variance of evidence, in each trial of the experiment, the presentation of the stimulus $S \in \{0, 1\}$ provides the observer with a sample of evidence x to select a response $R \in \{0, 1\}$. Participants are assumed to respond $R = 0$, when $x < \theta$, and $R = 1$ if $x > \theta$. θ is a free

parameter that represents the decision criterion. x is modeled as a random sample from a Gaussian distribution f :

$$x \sim f_{\mu_x, s^2} \quad (\text{A1})$$

The mean μ_x and variance s^2 of f are calculated as

$$\mu_x = (S - 0.5) \times d_m \quad (\text{A2})$$

$$s^2 = 1 + k \times d_m^2 \quad (\text{A3})$$

Thus, $s^2 = 1$ when $d_m = 0$, and the increase with d^2 is proportional to the parameter k for $d_m > 0$.

The decision variable y , which is evaluated when observers select one out of n confidence categories $C \in \{1, 2, \dots, n\}$, is again modeled as a random sample out of a Gaussian distribution:

$$y \sim f_{\mu_y, \sigma^2} \quad (\text{A4})$$

Its variance σ^2 represents imperfect access to the evidence underlying the decision, as well as, in the case of the WEV model, imperfect estimation of the quality of the stimulation. According to the WEV models, the mean of the decision variable y is given as

$$\mu_y = (1 - w) \times x + w \times (2R - 1) \times (d_m - \text{mean}(d)) \quad (\text{A5})$$

In this formula, w represents the relative weight that observers put on visibility or on the evidence for calculating their degree of confidence.

Participants are assumed to report a minimal degree of confidence $C = 1$ about the response $R = 0$ when $y > c_{01}$, and about a response $R = 1$ when $y < c_{11}$. The probability of a response $R = 0$ and confidence $C = 1$ can be calculated as:

$$P(R = 0 \wedge C = 1) = P(x < \theta) \times P(y > c_{01} | x) \quad (\text{A6})$$

$$P(R = 0 \wedge C = 1) = \int_{-\infty}^{\theta} f_{\mu_x, s^2}(x) \times \left(\int_{c_{01}}^{\infty} f_{\mu_y, \sigma^2} dy \right) dx \quad (\text{A7})$$

The probability of a response $R = 1$ and confidence $C = 1$ is calculated analogously:

$$P(R = 1 \wedge C = 1) = \int_{\theta}^{\infty} f_{\mu_x, s^2}(x) \times \left(\int_{-\infty}^{c_{11}} f_{\mu_y, \sigma^2} dy \right) dx \quad (\text{A8})$$

Participants report a medium degree of confidence $C = i$ given $R = 0$, if $c_{0(i+1)} < y < c_{0i}$, and likewise, a medium degree of confidence $C = i$ given $R = 1$, if $c_{1i} < y < c_{1(i+1)}$. The probability of confidence $C = i$ is given by:

$$P(R = 0 \wedge C = i) = \int_{-\infty}^{\theta} f_{\mu_x, s^2}(x) \times \left(\int_{c_{0(i+1)}}^{c_{0i}} f_{\mu_y, \sigma^2} dy \right) dx \quad (\text{A9})$$

$$P(R = 1 \wedge C = i) = \int_{\theta}^{\infty} f_{\mu_x, s^2}(x) \times \left(\int_{c_{1i}}^{c_{1(i+1)}} f_{\mu_y, \sigma^2} dy \right) dx \quad (\text{A10})$$

Finally, observers report a maximum degree of confidence $C = n$ about the response $R = 0$ when $y < c_{0(n-1)}$, and about the response $R = 1$ when $y > c_{1(n-1)}$. Therefore, the probability of confidence given $C = n$ is given by

$$P(R = 0 \wedge C = n) = \int_{-\infty}^{\theta} f_{\mu_x, s^2}(x) \times \left(\int_{-\infty}^{c_{0(n-1)}} f_{\mu_y, \sigma^2} dy \right) dx \quad (\text{A11})$$

$$P(R = 1 \wedge C = n) = \int_{\theta}^{\infty} f_{\mu_x, s^2}(x) \times \left(\int_{c_{1(n-1)}}^{\infty} f_{\mu_y, \sigma^2} dy \right) dx \quad (\text{A12})$$

The WEV model

The WEV model is identical to the WEV model with SOA-dependent variance, except that k is fixed at 0. Thus, in the WEV model, Equation A3 reduces to $s^2 = 1$.

The noisy SDT model

The noisy SDT model is identical to the WEV model with SOA-dependent variance, except that the parameters k and w are fixed at 0. Therefore, according to the noisy SDT model, Equation A3 is reduced to $s^2 = 1$, and Equation A5 to $\mu_y = x$.

The SDT model with SOA-dependent variance

According to the SDT model with SOA-dependent variance, the parameters σ and w are fixed at 0. At a result, the formulae in Equations A7 and A12 can be reduced to

$$P(R = 0 \wedge C = 1) = \int_{c_{01}}^{\theta} f_{\mu_x, s^2}(x) dx \quad (\text{A13})$$

$$P(R = 1 \wedge C = 1) = \int_{\theta}^{c_{11}} f_{\mu_x, s^2} dx \quad (\text{A14})$$

$$P(R = 0 \wedge C = i) = \int_{c_{0(i+1)}}^{c_{0i}} f_{\mu_x, s^2}(x) dx \quad (\text{A15})$$

$$P(R = 1 \wedge C = i) = \int_{c_{1i}}^{c_{1(i+1)}} f_{\mu_x, s^2} dx \quad (\text{A16})$$

$$P(R = 0 \wedge C = n) = \int_{-\infty}^{c_{0(n-1)}} f_{\mu_x, s^2}(x) dx \quad (\text{A17})$$

$$P(R = 1 \wedge C = n) = \int_{c_{1(n+1)}}^{\infty} f_{\mu_x, s^2} dx \quad (\text{A18})$$

The SDT model

According to SDT, not only are the parameters σ and w fixed at 0, but k is 0 as well. At a result, the formulae in Equations A13–A18 apply, except that s^2 is always equal to 1.

The postdecisional accumulation model

According to the postdecisional accumulation model, k is again fixed at 0. The formulae in Equations A7–A12 apply. However, μ_y is calculated in a different way than in the WEV model:

$$\mu_y = x + (2S-1) \times d_i \times b \quad (\text{A19})$$

The free parameter b represents the amount of postdecisional accumulation, expressed in terms of the sensitivity without postdecisional accumulation. Moreover, σ^2 is not a free parameter, as in the two WEV models or the noisy SDT model, but is always equal to b .

The two-channel model

According to the two-channel model, k is again fixed at 0. The formulae in Equations A7–A12 apply, except that σ^2 is not a free parameter, as in the two WEV models or the noisy SDT model, but is fixed at 1. Moreover, μ_y is calculated in a different way than in the WEV model:

$$\mu_y = a \times d_i \quad (\text{A20})$$

References

- Aitchison, L., Bang, D., Bahrami, B., & Latham, P. E. (2015). Doubly Bayesian analysis of confidence in perceptual decision-making. *PLoS Computational Biology*, 11, e1004519. <https://doi.org/10.1371/journal.pcbi.1004519>
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.
- Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics*, 55, 412–428.
- Barthelmé, S., & Mamassian, P. (2010). Flexible mechanisms underlie the evaluation of visual confidence. *Proceedings of the National Academy of Sciences*, 107, 20834–20839. <https://doi.org/10.1073/pnas.1007704107>
- Boldt, A., de Gardelle, V., & Yeung, N. (2017). The impact of evidence reliability on sensitivity and bias in decision confidence. *Journal of Experimental Psychology: Human Perception and Performance*, 43, 1520–1531. <https://doi.org/10.1037/xhp0000404>
- Boldt, A., & Yeung, N. (2015). Shared neural markers of decision confidence and error detection. *Journal of Neuroscience*, 35, 3478–3484. <https://doi.org/10.1523/JNEUROSCI.0797-14.2015>
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York, NY: Springer.
- Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, 6, 274–290. <https://doi.org/10.1177/1745691611406920>
- Dorfman, D. D., & Alf, E. (1969). Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals—Rating-method data. *Journal of Mathematical Psychology*, 6, 487–496. [https://doi.org/10.1016/0022-2496\(69\)90019-4](https://doi.org/10.1016/0022-2496(69)90019-4)

- Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision performance: A general Bayesian framework for metacognitive computation. *Psychological Review*, 124, 91–114. <https://doi.org/10.1037/rev0000045>
- Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B*, 367, 1338–1349. <https://doi.org/10.1098/rstb.2011.0417>
- Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, 62, 451–482. <https://doi.org/10.1146/annurev-psych-120709-145346>
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.
- Hebart, M. N., Schriever, Y., Donner, T. H., & Haynes, J. D. (2016). The relationship between perceptual decision variables and confidence in the human brain. *Cerebral Cortex*, 26, 118–130. <https://doi.org/10.1093/cercor/bhu181>
- Ince, D. C., Hatton, L., & Graham-Cumming, J. (2012). The case for open computer programs. *Nature*, 482, 485–488. <https://doi.org/10.1038/nature10836>
- Jang, Y., Wallsten, T. S., & Huber, D. E. (2012). A stochastic detection and retrieval model for the study of metacognition. *Psychological Review*, 119, 186–200. <https://doi.org/10.1037/a0025960>
- Kellen, D., & Klauer, K. C. (2015). Signal detection and threshold modeling of confidence-rating ROCs: A critical test with minimal assumptions. *Psychological Review*, 122, 542–557. <https://doi.org/10.1037/a0039251>
- Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B*, 367, 1322–1337. <https://doi.org/10.1098/rstb.2012.0037>
- Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, 455, 227–31. <https://doi.org/10.1038/nature07200>
- Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron*, 84, 1329–1342. <https://doi.org/10.1016/j.neuron.2014.12.015>
- Kyllingsbæk, S., & Bundesen, C. (2007). Parallel processing in a multifeature whole-report paradigm. *Journal of Experimental Psychology: Human Perception and Performance*, 33, 64–82. <https://doi.org/10.1037/0096-1523.33.1.64>
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Mahwah, NJ: Erlbaum.
- Maniscalco, B., & Lau, H. (2016). The signal processing architecture underlying subjective reports of sensory awareness. *Neuroscience of Consciousness*, 1:1–17. <https://doi.org/10.1093/nc/niw002>
- Marshall, L., & Bays, P. (2013). Obligatory encoding of task-irrelevant features depletes working memory resources. *Journal of Vision*, 12(9), 853. <https://doi.org/10.1167/12.9.853>
- 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. <https://doi.org/10.1016/j.cogpsych.2015.01.002>
- Morey, R. D., & Rouder, J. N. (2014). BayesFactor: Computation of Bayes factors for common designs (R package version 0.9.9). Retrieved from <http://cran.r-project.org/package=BayesFactor>
- Morin, A., Urban, J., Adams, P. D., Foster, I., Sali, A., Baker, D., & Sliz, P. (2012). Shining light into black boxes. *Science*, 336, 159–160. <https://doi.org/10.1126/science.1218263>
- Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, 7, 308–313.
- Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, 95, 109–133. <https://doi.org/10.1037/0033-2909.95.1.109>
- Paz, L., Insabato, A., Zylberberg, A., Deco, G., & Sigman, M. (2016). Confidence through consensus: a neural mechanism for uncertainty monitoring. *Scientific Reports*, 6, 21830. <https://doi.org/10.1038/srep21830>
- Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, 162, 8–13. <https://doi.org/10.1016/j.jneumeth.2006.11.017>
- Peirce, J. W. (2008). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, 2, 10. <https://doi.org/10.3389/neuro.11.010.2008>
- Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, 117, 864–901. <https://doi.org/10.1037/a0019737>
- Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience*, 19, 366–374. <https://doi.org/10.1038/nn.4240>
- R Core Team. (2014). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.r-project.org/
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85, 59–108. <https://doi.org/10.1037/0033-295X.85.2.59>
- Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, 20, 260–281. <https://doi.org/10.1016/j.tics.2016.01.007>
- Rausch, M., & Zehetleitner, M. (2014). A comparison between a visual analogue scale and a four point scale as measures of conscious experience of motion. *Consciousness and Cognition*, 28, 126–140. <https://doi.org/10.1016/j.concog.2014.06.012>
- Rausch, M., & Zehetleitner, M. (2016). Visibility is not equivalent to confidence in a low contrast orientation discrimination task. *Frontiers in Psychology*, 7, 591. <https://doi.org/10.3389/fpsyg.2016.00591>
- Rausch, M., & Zehetleitner, M. (2017). Should metacognition be measured by logistic regression? *Consciousness and Cognition*, 49, 291–312. <https://doi.org/10.1016/j.concog.2017.02.007>
- Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, 47, 877–903. <https://doi.org/10.1080/00273171.2012.734737>
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16, 225–237. <https://doi.org/10.3758/PBR.16.2.225>
- Sanders, J. I., Hangya, B., & Kepecs, A. (2016). Signatures of a statistical computation in the human sense of confidence. *Neuron*, 90, 499–506. <https://doi.org/10.1016/j.neuron.2016.03.025>
- Schwarz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics*, 6, 461–464. <https://doi.org/10.1214/aos/1176348654>
- Spence, M. L., Dux, P. E., & Arnold, D. H. (2015). Computations underlying confidence in visual perception. *Journal of Experimental Psychology: Human Perception and Performance*, 42, 671–682. <https://doi.org/10.1037/xhp0000179>
- Tversky, A., & Kahneman, D. (1975). Judgment under uncertainty: Heuristics and biases. In *Utility, probability, and human decision making* (pp. 141–162). Amsterdam, The Netherlands: Springer Netherlands.
- Urai, A. E., Braun, A., & Donner, T. H. (2017). Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nature Communications*, 8, 14637. <https://doi.org/10.1038/ncomms14637>
- van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. *eLife*, 5, e12192. <https://doi.org/10.7554/eLife.12192>

- van den Berg, R., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). Confidence is the bridge between multi-stage decisions. *Current Biology*, 26, 3157–3168. <https://doi.org/10.1016/j.cub.2016.10.021>
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7, 632–638. <https://doi.org/10.1177/1745691612463078>
- Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science*, 6, 291–298. <https://doi.org/10.1177/1745691611406923>
- Wickens, T. D. (2002). *Elementary signal detection theory*. New York, NY: Oxford University Press.
- Xu, Y. (2010). The neural fate of task-irrelevant features in object-based processing. *Journal of Neuroscience*, 30, 14020–14028. <https://doi.org/10.1523/JNEUROSCI.3011-10.2010>
- Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B*, 367, 1310–1321. <https://doi.org/10.1098/rstb.2011.0416>
- Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience*, 6, 79. <https://doi.org/10.3389/fnint.2012.00079>