# The Precision Test of metacognitive sensitivity and confidence criteria

Derek H. Arnold[1], Mitchell Clendinen[1], Alan Johnston[2], Alan L.F. Lee[3] & Kielan Yarrow[4]

1. School of Psychology, The University of Queensland | 2. School of Psychology, The University of Nottingham | 3. Department of Psychology, Lingnan University, Hong Kong | 4. School of Psychology, City University London

**Humans experience feelings of confidence in their decisions. In perception, these feelings are typically accurate – we tend to feel more confident about correct decisions. The degree of insight people have into the accuracy of their decisions is known as metacognitive sensitivity. Currently popular methods of estimating metacognitive sensitivity are subject to interpretive ambiguities because they assume people have normally shaped distributions of different experiences when they are repeatedly exposed to a single input. If this normality assumption is violated, calculations can erroneously underestimate metacognitive sensitivity. Here, we describe a means of estimating metacognitive sensitivity that is more robust to violations of the normality assumption. This improved method can easily be added to standard behavioral experiments, and the authors provide Matlab code to help researchers implement these analyses and experimental procedures.**

**Key Words:** Perceptual metacognition; Confidence; Signal Detection Theory

**Public Significance Statement:** Signal-detection theory is one of the most popular frameworks for analysing data from experiments of human behaviour – including investigations of confidence. The authors demonstrate that if a key assumption of this framework is violated, analyses can lead to unwarranted conclusions. They develop a new and more robust measure of confidence.

**Correspondence to:** d.arnold@psy.uq.edu.au

**Acknowledgements:** This research was supported by a Discovery Project Grant DP200102227, funded by the Australian Research Council, awarded to D.H.A.

**Conflict of interest:** The authors declare no competing financial interests.

**Data and materials availability:** All data and analysis scripts for this project will be made available via UQeSpace https://espace.library.uq.edu.au

**Funding:** This research was supported by an ARC Discovery Project Grant awarded to DHA.

## Introduction

People experience levels of confidence when making decisions (Fleming et al., 2012; Yeung & Summerfield, 2012). In studies of perception, we can verify that these feelings are typically accurate, with higher levels of confidence experienced for better decisions (Keane et al., 2015; Li et al., 2014; Peters et al., 2017). This is indicative of a level of insight into the quality of our decisions – known as metacognitive sensitivity (Maniscalco & Lau, 2012, 2016). When confidence appears to have been informed by the same information as our decisions, people are said to be metacognitively ideal (Fleming and Lau, 2014; Maniscalco & Lau, 2012, 2016). However, when confidence appears to have been shaped by different information people are said to be metacognitively insensitive (Fleming and Lau, 2014; Maniscalco & Lau, 2012, 2016).

A range of methods have been used to estimate metacognitive sensitivity (or rather the degree of metacognitive insensitivity). A previously popular approach was to assess correlations between the accuracy of perceptual decisions and confidence ratings (e.g. Nelson, 1984). This approach, however, had a serious shortcoming. Such measures can confound confidence bias – an individuals' propensity to report that they are confident, with the desired measure of metacognitive sensitivity. Two people could have equal levels of metacognitive sensitivity, have equally heightened feelings of confidence when they are making a well-informed relative to an ill-informed decision, but have different levels of willingness to report that they are confident. This confidence bias can obscure an individuals' true level of metacognitive sensitivity (see Masson & Rotello, 2009).

*Currently popular SDT-based analyses of confidence*

More recent attempts to measure metacognitive sensitivity have relied on an extrapolation of the Signal-Detection-Theory (SDT) framework (Fleming & Lau, 2014; Maniscalco & Lau, 2012). In addition to the standard SDT framework, these approaches assume people use criteria to demarcate when a decision has been regarded as having evoked a relatively low or a relatively high level of confidence (see Figure 1). With this added assumption, two sets of estimates of perceptual bias and sensitivity can be calculated from a common dataset. Traditional estimates of sensitivity (d') and bias can be calculated from perceptual category decisions (Green & Swets, 1966). Researchers can then also backwards infer a second estimate of sensitivity and bias, from analyses that are informed by the proportions of correct (e.g. Figure 1, bottom panel, light red and blue shaded regions) and incorrect (e.g. Figure 1, bottom panel, dark purple shaded regions) high-confidence decisions (Fleming & Lau, 2014; Maniscalco & Lau, 2012). The key statistic this 'type-2 confidence analysis' delivers is meta d', which is regarded as an estimate of metacognitive sensitivity, which is expressed relative to d' estimates by either calculating difference scores (meta d' – d') or ratios (meta d' : d'). The standard finding is that meta d' estimates fall short of d' estimates (e.g. Maniscalco et al., 2016; Rausch et al., 2015), and this is regarded as evidence that confidence ratings have been shaped by a different source of information or noise relative to perceptual decisions.

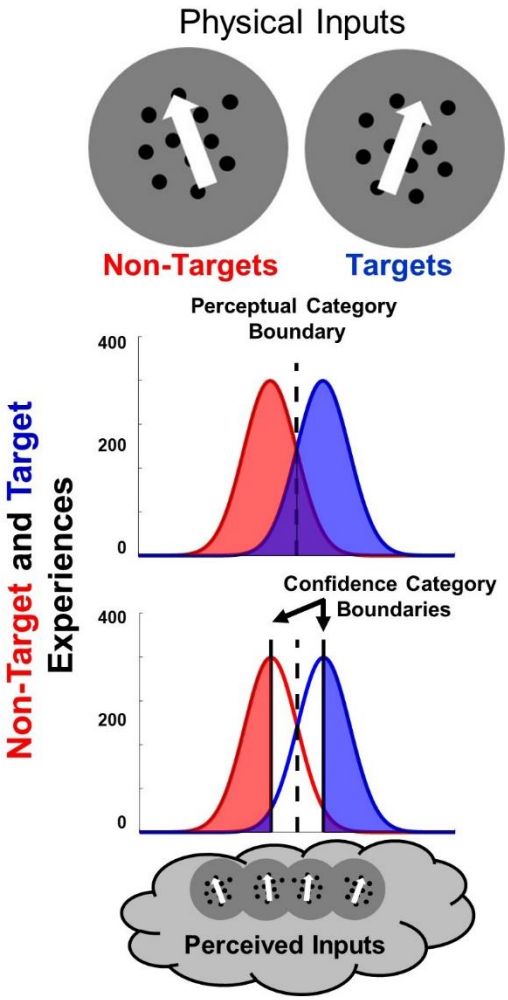


**Figure 1.** Graphic describing assumptions underlying popular extensions of the SDT framework, used to analyze confidence. **Above**, a standard SDT decision space, with distributions describing different experiences resulting from repeated exposures to a non-target (red) and a target (blue) input. These are thought to be referenced against a stable perceptual category boundary (bold black dotted vertical line) when people make categorical decisions (i.e. are test dots moving toward the left or right of vertical). **Below**, People are also believed to reference their perceptual experiences against stable confidence criterion values (bold black vertical lines) when deciding if they should endorse a perceptual decision with low (values toward the center from the two bold black lines) or high confidence.

SDT-based analyses of confidence are regarded as an improvement on correlations of confidence with decisional accuracy (e.g. Nelson, 1984), as they provide independent estimates of metacognitive sensitivity and confidence bias (Fleming & Lau, 2014; Maniscalco & Lau, 2012). These analyses can, however, encourage erroneous conclusions if they are inadvertently informed by experiential distributions (E.D.s) that are non-normally

shaped. An E.D. describes the different experiences a human will have following repeated exposures to a common physical input. If these are non-normally shaped, either because the E.D. is skewed (see Arnold et al., 2023), or because the E.D. has fatter tails (i.e. a greater number of extreme experiences, described as excess kurtosis) than a normally-shaped distribution, SDT-based analyses can promote erroneous conclusions (see Arnold et al., 2023; Miyoshi et al., 2022). This is a serious consideration, as there is good evidence that E.D.s in visual perception are both subject to excess kurtosis (e.g. Acerbi et al., 2012; Anderson, 2014; Bays, 2016; Jabar & Anderson, 2015) and to localized skews (e.g. Appelle, 1972; Girshick et al., 2011; Storrs & Arnold., 2015b).

*Evidence for non-normally shaped psychological dimensions in perception*

The precision of perceptual experiences is not uniform across many visual dimensions. Humans are, for instance, more sensitive to direction and to orientation differences around cardinal (vertical and horizontal) as opposed to oblique angles (Appelle, 1972; Dakin et al., 2005; Girshick et al., 2011; Storrs & Arnold, 2015b). Moreover, spatial acuity scales with distance from fixation (Pollack & Mueller, 2005), and people are more sensitive to differences between slower than between faster speeds (e.g. Stocker & Simoncelli, 2006). In each of these cases, experiences triggered by repeated exposure to a common input will likely be associated with a greater range of different experiences that extend into regions of a psychological dimension characterized by *less* sensitivity/precision, and with a smaller range of different experiences that extend into regions of the dimension that are associated with greater sensitivity/precision. In sum, this would produce a localized skew of the given E.D. So, in human vision skewed E.D.s are likely more common than non-skewed dimensions.

There is also good evidence that human E.D.s are characterized by excess kurtosis. A number of studies have estimated the shape of E.D.s by having participants reproduce their perceptual experiences (e.g. of perceived tilts, or interval durations). These studies have suggested that human E.D.s are characterized by excess kurtosis (Acerbi et al., 2012; Anderson, 2014; Bays, 2016; Jabar & Anderson, 2015). In addition, the distribution of the responses of tuned cells (i.e. the response rates of cells that are tuned to different orientations) is often marked by excess kurtosis, particularly when the cells are tuned for inputs that have a learnt relevance (e.g. Failor et al., 2021) and when inputs are natural images (e.g. Field, 1994; Olshausen and Field, 1996). Again, this evidence suggests that non-normally shaped psychological dimensions might dominate human visual perception, as opposed to normally shaped dimensions.

*Experimenters cannot easily detect non-normally shaped distributions that can undermine analyses of confidence*

The possibility that a psychological dimension might be non-normally shaped would not be an issue if experimenters had an easy means of estimating the precise shape of the E.D.s evoked by repeated exposures to given test inputs. However, it has been established that current means of assessing if E.D.s might have a non-normal shape are insensitive to deviations from normality that are sufficient to cause meta d' to be systematically underestimated (see Arnold et al., 2023). So, there is an interpretive ambiguity. Any empirical finding suggestive of a meta d' : d' difference could either be due to confidence having been informed by different information relative to perceptual decisions, or it could have arisen because perceptual decisions have been informed by a common set of non-normally shaped E.D.s.

*A new Precision Test of metacognitive sensitivity and confidence criteria*

Given the conceptual limitations of currently popular means of estimating metacognitive sensitivity and confidence criteria, we were motivated to devise a more robust method. Our approach is based on an experimental design, wherein people must decide if inputs belong to one of two categories (a binary forced choice task, a standard procedure in perception studies). We estimate the precision with which participants transition from making relatively low- to relatively high-confidence perceptual category decisions, and compare these estimates to the precision with which they transition between predominantly making different types of perceptual category decision (regardless of confidence, see Figure 2). We fit cumulative Gaussian functions to data, and take the standard deviations of function fits as a metric of the precision of the different types of judgment (about perception and confidence). Additionally, the horizontal separation of distributions describing transitions in confidence, from a central distribution describing the transition between making different types of perceptual category decision, provides an estimate of the participants' criterion for reporting that they had a relatively high-level of perceptual confidence (see Figure 2).

We control for the impact of individual differences in confidence bias by categorizing confidence ratings as low or as high relative to each individual's average confidence rating across an experimental session. This can be done for confidence ratings taken from settings made along a continuous scale, or for confidence ratings selected from a limited range of Likert scale options.

Some readers might regard our treatment of confidence data (where we transform an individual's confidence ratings into a dichotomous, high / low measure) as a weakness. We

disagree. First, we would point out that our treatment of confidence data does not require that researchers disregard any greater granularity of confidence reports that they might have recorded during an experiment. They can use confidence data in additional analyses (relative to the calculations that inform our analyses). Second, we would remind readers that our transformation of confidence data into a dichotomous measure serves to control for individual differences in confidence bias, which is a strength. Finally, the benefits of having a large number of different levels of reported confidence are uncertain. Each level of reported confidence will be subject to criteria (to which participants must relate their feelings of confidence to decide what rating should be used to describe their feeling of confidence). So, it is probable that allowing participants to report on a greater number of levels of confidence will introduce more criterion noise. These issues are often neglected by researchers. In sum, we do not accept that the dichotomous treatment of confidence data is necessarily a weakness.

From our testing procedure, two functions result that describe transitions in confidence. The first (see blue data, Figure 2) plots the proportion of trials on which a participant has either made a Category A decision with low confidence (e.g. a low-confidence left tilt decision), or any Category B decision (e.g. a right tilt decision, regardless of confidence). The second (see red data, Figure 2) plots the proportion of trials on which a participant has endorsed a Category B perceptual decision with high confidence.

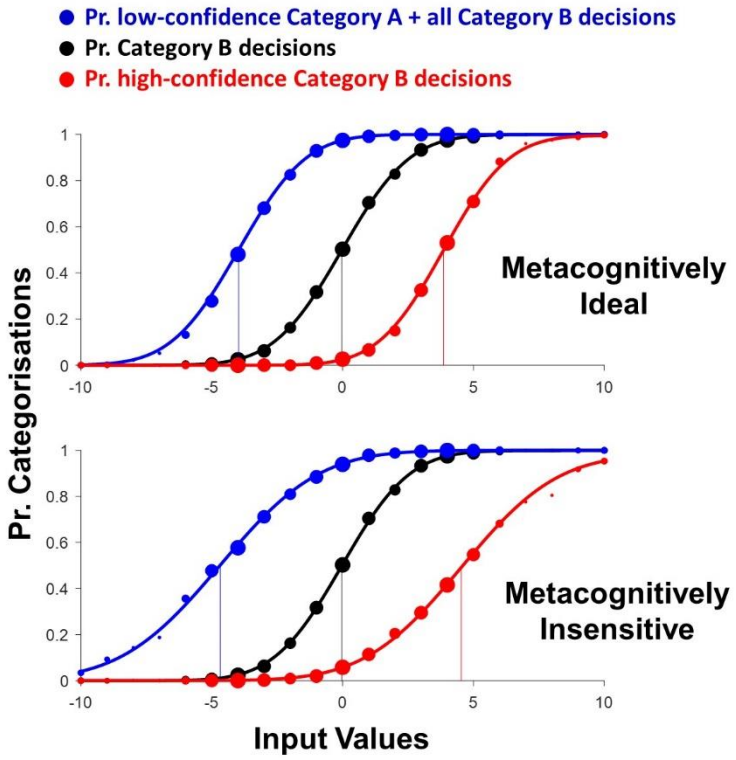


**Figure 2.** Cumulative gaussian functions fit to **1)** proportions of simulated trials resulting in either a low-confidence category A perceptual decision or any category B decision (**blue data**), **2)** proportions of simulated trials resulting in any category B perceptual decision (**black data**), or **3)** proportions of trials resulting in high-confidence category B decisions (**red data**). The average standard deviations of **blue** and **red** functions are regarded as estimates of the precision of confidence judgments, and the standard deviation of **black** functions is regarded as an estimate of the precision of categorical perceptual decisions. A metacognitively ideal (top) and a metacognitively insensitive (bottom) participant are simulated. The standard deviations of all functions are inversely related to the noise impacting the decisions described by the functions.

If confidence ratings are informed by the same information as perceptual category decisions, and the criteria people use to categorize their feelings of confidence are equally stable relative to the criteria people use to make perceptual category decisions, then estimates of the precision of confidence ratings and of the precision of perceptual category decisions should be equal (i.e. functions fitted to data describing these decisions should have a common standard deviation and maximal slope). This situation can be described as metacognitively ideal, as the same information has informed both the perceptual decisions and confidence ratings (see Figure 2, upper panel). If, however, confidence ratings are shaped by different

information (either from processes that generate our feelings of confidence, or from processes that are used when applying confidence criteria), then estimates of the precision of confidence ratings can fall short of estimates of the precision of perceptual decisions (i.e. functions fit to data describing confidence decisions will have a larger standard deviation and shallower maximal slope than functions fit to data describing perceptual category decisions). This situation provides evidence for metacognitive insensitivity (see Figure 2, lower panel).

*The current study*

Having described our new Precision Test to measure metacognitive sensitivity and confidence criteria, we now turn to considering if it is a conceptual improvement on what is currently the most popular SDT-based analysis of confidence (Maniscalco & Lau, 2012). We establish this in three sets of simulations. In the first set, we assess how the two approaches are impacted by distributions that have different skews. In a second set of simulations, we assess how the two approaches are impacted by distributions that have different levels of kurtosis. Across both these sets of simulations, we show that our new Precision Test is more robust to violations of the normality assumption. In a third set of simulations we show that this improvement is achieved by **1)** implementing a dynamic sampling routine, that concentrates testing at three important points of a psychological dimension – the two points of transition between making low and high confidence decisions (see the blue and red functions, Figure 2), and the point at which people transition between making different category decisions regardless of confidence (see the black function, Figure 2), and **2)** by avoiding use of a bespoke maximum likelihood estimation routine, that exaggerates the impact of non-normally shaped psychological dimensions.

Having established that our Precision Test is a conceptual improvement, we move onto demonstrating that our approach can be implemented in behavioural experiments with human participants. We report on 3 experiments. The first two re-examine the impact of the range of direction signals on the precision of judgments concerning motion direction perception and confidence (Spence et al., 2016). The third experiment re-examines the impact of confirmation bias on the precision of perceptual decisions and confidence (e.g. Braun et al., 2018; Rollwage et al., 2020).

**Simulation 1a: The Precision Test of metacognitive sensitivity and distribution skews**

We chose to develop matlab scripts to control and analyze the results of simulated experimental sessions, mimicking a study of tilt perception. This was not conceptually necessary, as we could have simply generated distributions with specified characteristics (as we do in Simulation 3). However, we wanted to develop and test matlab scripts to control the dynamic sampling of appropriate test inputs, which we could then use in future experiments with human participants (see Experiments 1 – 3). The Matlab code we used to implement this set of simulations is provided as Supplemental Code #1. In supplemental material we also provide a tutorial describing how our code controls the sampling of test values, and how it can be used to analyze the results of an experimental session. Also see our webpage: Metacognitive Sensitivity: Precision Test (uq.edu.au). Here we provide a less detailed description of these procedures, sufficient for readers to understand the conceptual implications of this set of simulations.

All simulations are of a forced-choice binary perceptual categorization task (left / right tilt). Each simulated trial results in a value being sampled from a distribution that has been pre-

generated using the Matlab ‘pearsrnd’ command. This allows us to specify the mean, S.D. (2), kurtosis (3, so an excess kurtosis of 0) and skew of distributions. This command also allows us to simulate sampling noise (as the command randomly samples a specified number of values from an underlying distribution). These pre-generated distributions mimic the E.D.s that would be experienced by human participants if they were repeatedly exposed to a given test input (see Figure 1). We created distributions for each of a range of potential test values (that could be sampled by our adaptive staircase procedures). These ranged from a mean test value of -5 S.D. units to +5 S.D.s (in steps of 0.5 S.D.s).

Our code runs 5 interleaved staircase procedures, designed to target sampling at **1)** the two asymptotes of a psychological dimension (where categorical responding nears floor and ceiling), **2)** the two points of transition between making low and high confidence perceptual category decisions, and **3)** the perceptual category boundary (i.e. subjective vertical). Test values on simulated trials are categorized as having been left (values <0) or right (values >=0) tilted. For confidence ratings the sign (+/-) of test values is discarded, and the value is classified as having elicited a relatively low or a relatively high level of confidence (see Figure 1).

Different levels of skew difference are sampled across simulated experimental sessions. Distributions informing analyses are either associated with 0 skew (i.e. by normally shaped distributions), or with distributions of negatively signed test values that are associated with a positive skew (of 0.16 or 0.32) and with distributions of positively signed test values that are associated with distributions with negative skews (of 0.16, or 0.32). So, Simulation 1a samples skew differences of 0, 0.32 and 0.64.

Sampled test values are referenced against a perceptual category criterion (of 0, to simulate unbiased participants). For each level of differential skew, we simulate participants with a range of 81 different high confidence criterion values (ranging from 1 to 5 S.D. units, in 0.05 S.D. steps). These high confidence criterion values represent the absolute difference in orientation (+/- from the 0° perceptual category boundary) required to elicit a relatively high level of decisional confidence. Lesser differences evoke a relatively low level of decisional confidence.

Half our simulations are of metacognitively ideal participants, who have no additional noise associated with their confidence ratings. Other simulations are of metacognitively insensitive participants, whose confidence ratings are subject to an additional source of noise. This is achieved by randomly varying the confidence criterion by +/- 1.25 S.D. units from its nominal value on each simulated trial.

Each experimental session in simulations 1a and 1b generates a set of 3 distributions (as depicted in Figure 2). These are analyzed as per descriptions in the section subtitled ‘A new Precision Test of metacognitive sensitivity and confidence criteria’ (also see Supplemental material). So, simulation 1a is of 243 participants (3x different skew levels x 81 high confidence criteria) who are metacognitively ideal (see Figure 3a), and of 243 participants who are metacognitively insensitive (see Figure 3b).

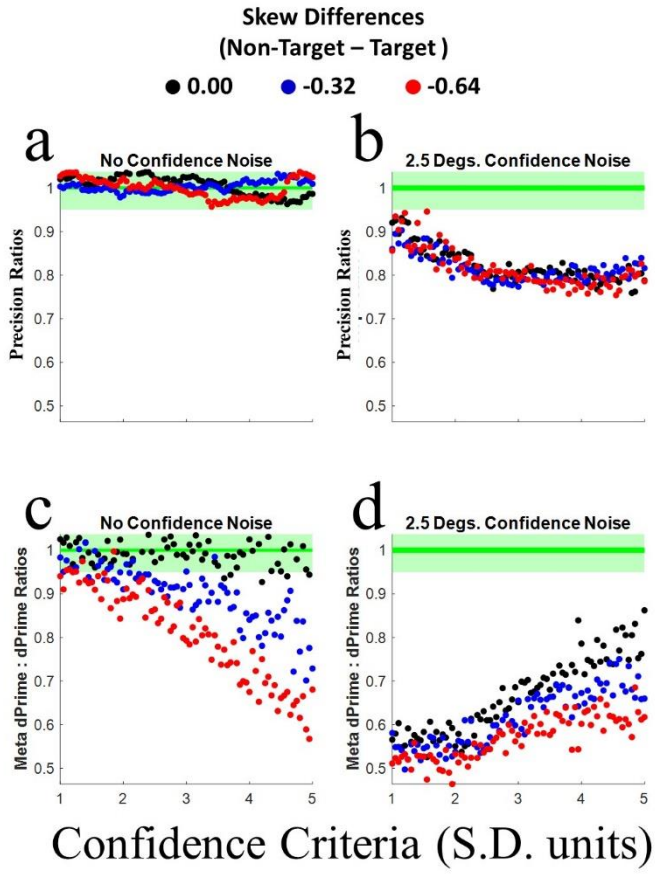


**Figure 3. a)** Scatterplot of confidence : perceptual precision ratios (Y axis) and confidence criterion test values (X axis). Data relate to metacognitively ideal simulated participants. The bold horizontal green line depicts ratios of 1, and green shaded region depict ratios +/- 5%. Black data points depict ratios calculated from normally-shaped E.D.s, blue data depict ratios calculated from E.D.s with a skew difference of 0.32, and red data depict ratios calculated from E.D.s with a skew difference of 0.64 (see main text for a further explanation). **b)** Details are as for Figure 3a, but these data relate to metacognitively insensitive simulated participants – due to trial-by-trial confidence criterion noise. **c)** Details are as for Figure 3a, but these data relate to meta d' : d' ratios (Y axis) calculated for metacognitively ideal simulated participants. **d)** Details are as for Figure 3c, but these data relate to metacognitively insensitive simulated participants – due to trial-by-trial confidence criterion noise.

**Results: Simulation 1a**

Confidence : perceptual sensitivity ratios are plotted in Figure 3a, for simulations of metacognitively ideal participants. As the simulated participants were metacognitively ideal (i.e. the same information and noise has informed both the simulated perceptual decisions and confidence ratings), these ratios should cluster about a value of 1. As can be seen in Figure 3a, this is a reasonable description of these data, with all ratios falling within 5% of this ratio.

These data show that our Precision Test of metacognitive sensitivity and confidence criteria is robust when analyses are informed by distributions that are non-normally shaped due to having different skews.

Confidence : perceptual precision ratios are plotted in Figure 3b, for metacognitively insensitive simulated participants (due to added confidence criterion noise). For these simulated participants ratios that describe the precision of confidence ratings, relative to the precision of perceptual judgments, should be less than 1. Note that all the simulated participants are correctly identified as metacognitively insensitive, as all ratios are >5% below a ratio of 1.

**Simulation 1b: SDT-based analyses of confidence and skew**

Details for Simulation 1b are as for Simulation 1a, with the following exceptions. Just 2 test inputs are simulated, with each associated with a distribution of 20,000 values generated using the Matlab 'pearsrnd' command. The two distributions had mean values of -0.5 S.D. units (to simulate an E.D. from repeated exposures to a Non-Target, left tilted input, see Figure 1) and +0.5 S.D. units (to simulate an E.D. from repeated exposures to a Target, right tilted input). All other distribution details are as described for Simulation 1a.

Each E.D. value was categorized as either having elicited a high-confidence left tilt decision, a low-confidence left tilt decision, a low-confidence right tilt decision, or a high-confidence right tilt decision – with confidence ratings evaluated against the same high confidence criterion values as described for Simulation 1a. This enabled us to implement a popular SDT-based analysis of confidence (Maniscalco & Lau, 2012) for each simulated participant. This

delivered two sets of 243 d' and meta d' estimates, one for metacognitively ideal simulated participants, and one for simulated metacognitively insensitive simulated participants.

**Results: Simulation 1b**

Meta d' : d' ratios are plotted in Figure 3c, for analyses of data from simulated participants who were metacognitively ideal. As can be seen in Figure 3c, these are only broadly clustered about a ratio of 1 when the underlying E.D.s are normally shaped (black data). When analyses are instead informed by differently skewed E.D.s, nearly all ratios were > 5% below a ratio of 1, and this effect scaled with the magnitude of the confidence criterion values. These data show that a currently popular S.D.T. based analysis of confidence (Maniscalco & Lau, 2012) can encourage erroneous conclusions when analyses are informed by distributions that have different skews. Here the same information and noise has informed d' and meta d' calculations, and yet the results would encourage a researcher to believe the simulated participants had been metacognitively insensitive (i.e. that a different source of information or noise had shaped confidence ratings).

Meta d' : d' ratios are plotted in Figure 3d, for analyses of data from simulated participants who were metacognitively insensitive. As can be seen in Figure 3d, meta d' : d' ratios are all substantially less than 1, and this effect is negatively scaled with the magnitude of confidence criterion values (see Figure 3d). While these data show that metacognitively insensitive simulated participants can be identified by a SDT based analysis of confidence (Maniscalco & Lau, 2012), these classifications are not diagnostic, as the same analyses have suggested that many metacognitively ideal simulated participants are also metacognitively insensitive (see Figure 3c).

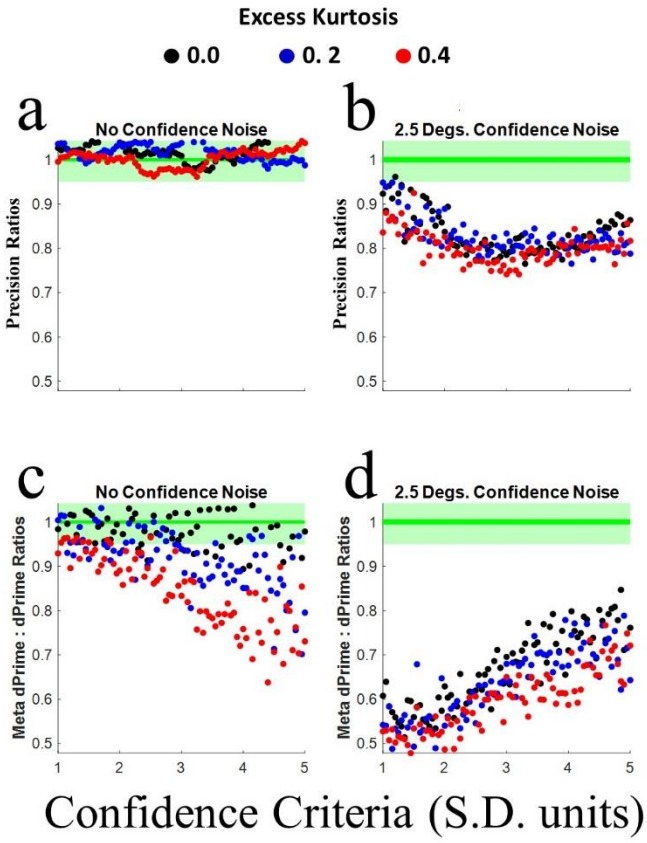


**Figure 4. a)** Scatterplot of confidence : perceptual standard deviation ratios (Y axis) as a function of confidence criterion values (X axis). These data relate to metacognitively ideal simulated participants. The bold horizontal green line depicts ratios of 1. The green shaded region depicts ratios between 0.95 and 1.05. Black data points depict ratios calculated from normally-shaped E.D.s, blue data points depict ratios calculated from E.D.s with an excess kurtosis of 0.2, and red data points depict ratios calculated from sessions that sampled E.D.s with an excess kurtosis of 0.4 **b)** Details are as for Figure 4a, but these data relate to metacognitively insensitive simulated participants – due to trial-by-trial confidence criterion noise. **c)** Details are as for Figure 4a, but these data relate to meta d' : d' ratios (Y axis) calculated for metacognitively ideal simulated participants. **d)** Details are as for Figure 4c, but these data relate to metacognitively insensitive simulated participants – due to trial-by-trial confidence criterion noise.

**Simulation 2a: The Precision Test of metacognitive sensitivity and excess kurtosis**

Details for Simulation 2a are as for Simulation 1a, with the following exceptions.

Code used to run this set of simulations has been provided as Supplemental Code #2. In this particular set, we simulate participants who have E.D.s that are characterized by different

levels of kurtosis. All E.D.s were set to 0 skew, and they were either normally shaped (with an excess kurtosis of 0), or they had excess kurtosis levels (fatter-tailed E.D.s) of 0.2 or of 0.4.

**Results: Simulation 2a**

For metacognitively ideal simulated participants, ratios describing the precision of confidence judgments : the precision of perceptual judgments are plotted in Figure 4a. As these simulated participants were metacognitively ideal, these ratios should have a constant value of approximately 1. This was true for all sampled high confidence criterion values, and for each level of excess kurtosis (0 – black datapoints, 0.2 – blue data points, and 0.4 – red data points). These data show that our Precision Test is robust when analyses are informed by E.D.s that have different levels of kurtosis.

Data plotted in Figure 4b are similar to data plotted in Figure 4a, but these data were calculated for simulated participants who were metacognitively insensitive. These ratios should therefore be <1. As can be seen in Figure 4b, nearly all ratios fall >5% below a ratio of 1 (with this effect initially increasing with confidence criterion values, before plateauing for confidence criterion values > 2 S.D.s). These data show that our Precision Test can detect genuine cases of metacognitive insensitivity, caused by confidence ratings being shaped by an additional source of information or noise relative to perceptual decisions.

**Simulation 2b: SDT-based analyses of confidence and excess kurtosis**

Details for Simulation 2b are as for Simulation 1b, with the following exceptions. All Target and Non-Target distributions had 0 skew. Instead, distributions had different levels of

kurtosis. Simulated E.D.s were either normally shaped (with an excess kurtosis of 0), or they have excess kurtosis levels of 0.2 or of 0.4.

**Results: Simulation 2b**

Meta d' : d' ratios are plotted in Figure 4c, for experimental sessions that simulated participants who were metacognitively ideal. As these simulated participants were metacognitively ideal (i.e. perceptual decisions and confidence ratings were generated from a common dataset, subject to the same sampling noise), meta d' : d' ratios should cluster about a value of 1. As can be seen in Figure 4c, this was only true in a broad sense when the simulated E.D.s were normally shaped (black data). When analyses were instead informed by E.D.s that had excess kurtosis (of 0.2 – blue data, or of 0.4 – red data) almost all ratios were > 5% below a ratio of 1, and this effect scaled with confidence criterion values (see Figure 4c). These data show that a SDT based analysis of confidence (Maniscalco & Lau, 2012) can encourage erroneous conclusions when analyses are informed by E.D.s that are non-normally shaped due to having excess kurtosis. Here the same data has informed d' and meta d' calculations, and yet results would encourage researchers to conclude that the simulated participants had been metacognitively insensitive.

Meta d' : d' ratios are plotted in Figure 4d, for simulated participants who were metacognitively insensitive. As can be seen in Figure 4d, meta d' : d' ratios are all substantially less than 1, and this effect is negatively scaled with confidence criterion values (see Figure 4d). While these data once again show that simulated participants would be correctly identified by a SDT based analysis of confidence (Maniscalco & Lau, 2012) as metacognitively insensitive, these classifications would not be diagnostic as the same

analyses have suggested many metacognitively ideal simulated participants were also metacognitively insensitive (see Figure 4c).

**Simulation 3: Why is our Precision Test of metacognitive sensitivity and confidence criteria more robust to violations of the normality assumption?**

Our new Precision Test is alike SDT based analyses of confidence (Fleming & Lau, 2014; Maniscalco & Lau, 2012), in that both approaches fit functions to data (in our case, cumulative Gaussian functions) that assume the underlying E.D.s are normally shaped (for a discussion of these issues, see Yarrow et al., 2011). So, why are our analyses more robust to violations of this shared assumption.

There are a two main differences between our new Precision Test and many SDT based studies of confidence. **1)** The main difference is that our procedures test many different test values, whereas many SDT based studies only sample 2 test values (e.g. Samaha & Denison, 2022; Mazor et al., 2020). This is orchestrated via procedures that target sampling at 5 points of an individual's psychological dimension – the two asymptotes of the psychological dimension (where categorical responding nears floor and ceiling), the two points of transition between low and high confidence categorical decisions (see blue and red functions in Figure 2), and the perceptual category boundary (see the black function in Figure 2). **2)** In contrast to our approach, SDT based analyses of confidence (e.g. Fleming & Lau, 2014; Maniscalco & Lau, 2012) use a bespoke maximum likelihood estimation procedure to arrive at estimates of metacognitive sensitivity. In Simulation 3, we show that both of these differences combine to ensure that our Precision Test is more robust to violations of the normality assumption.

Code for these simulations is available as Supplemental Code #3. In each simulation of this set, we sample 3 levels of excess kurtosis – with values of 0 (normally shaped distributions, black data), 0.2 and 0.4. For these simulations we use the matlab command fitglm.m, as this enables us to fit functions to just two data points – which allows us to compare our precision based test to SDT based analyses of confidence, while matching for the number of datapoints that have informed each analysis.

In our first set of simulations, we use a SDT based analysis of confidence (Maniscalco & Lau, 2012) to assess metacognitive sensitivity (meta d': d' ratios). Details for these analyses are largely similar to analyses depicted in Figure 4c (i.e. simulated participants are metacognitively ideal). Here the means of the E.D.s are fixed (+/- 2 S.D.s from the 0 perceptual category boundary) for all simulated participants – regardless of confidence criteria. These data serve to illustrate, once more, that a popular SDT based analysis of confidence can create a false impression, that metacognitively ideal participants are metacognitively insensitive (see Figure 5a).

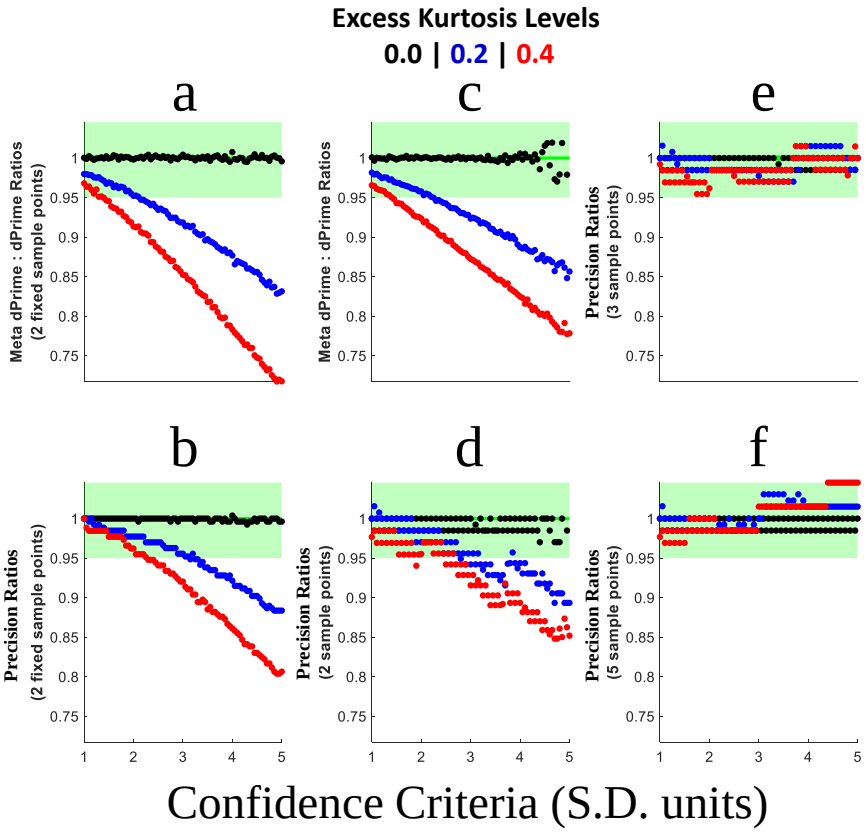


**Figure 5. a)** Scatterplot of meta d': d' ratios (Y axis) as a function of confidence criterion values (X axis). These data relate to metacognitively ideal simulated participants, and are calculated from E.D.s with 3 levels of excess kurtosis, 0 (normal distributions), 0.2 and 0,4. The mean values of the two distributions are set to +/- 0.5 S.D.s. The shaded green region depicts +/- 5% from the target ratio (1). See main text for further details. **b)** Details are as for Figure 5a, but data relate to Confidence Precision : Perceptual Precision ratios (Y-axis) calculated from functions fit to just two datapoints, to match the analyses that have informed the SDT based analyses of confidence depicted in Figure 5a. **c-d)** Details are as for Figure 5a-b, but data relate to analyses informed by sets of 2 data points calculated from distributions with mean values that are set to the confidence criteria used for each analysis. **e)** Details are as for Figure 5d, but data relate to Confidence Precision : Perceptual Precision ratios (Y-axis) calculated from functions fit to three datapoints, calculated from distributions with mean values set to the 2 confidence criteria used for each analysis and to the perceptual category boundary (0). **f)** Details are as for Figure 5e, but data relate to Confidence Precision : Perceptual Precision ratios (Y-axis) calculated from functions fit to five datapoints, inclusive of those described for Figure 5e and 2 distributions with mean values set to 1.5 x the confidence criteria used for each analysis.

In our second set of simulations, we use our Precision Test procedure to assess metacognitive sensitivity (confidence precision : perceptual precision ratios) based on estimates derived from functions fit to just two datapoints (derived from the same data that has informed the SDT based analyses depicted in Figure 5a). These data illustrate two important things. First,

our Precision Test can also give rise to a false impression that metacognitively ideal participants are metacognitively insensitive, when function fits are informed by *just two* datapoints. Second, the degree to which our testing procedure is undermined by excess kurtosis is reduced relative to a popular SDT based analysis of confidence (Maniscalco & Lau, 2012). Note that in Figure 5a, meta d' : d' ratios reach a minimum < 0.75, whereas in Figure 5b confidence precision : perceptual precision ratios reach a minimum >0.8. These two sets of analyses were matched, in terms of the number of datapoints (2) informing each individual analysis. The only substantive difference was that the SDT based analyses involves a bespoke maximum likelihood estimation routine to arrive at estimates of metacognitive sensitivity – so this procedure is evidently more susceptible to violations of the normality assumption (in the form of excess kurtosis) than our Precision Test.

In the next 2 sets of simulations, we repeat analyses depicted in Figure 5a and in 5b, but set the mean value of E.D.s to the 2 confidence criterion values that inform each analysis (e.g. to +/- 2 S.D.s when the confidence criterion is 2 S.D.s). This marginally improves both sets of analyses, such that minimum ratios achieved in both sets of analyses are now greater (closer to the target ratio of 1). However, it is also clear this has been insufficient to prevent a false impression, that metacognitively ideal participants are metacognitively insensitive, being promoted by both sets of analyses.

Our final two sets of simulations implement our Precision Test procedure, to estimate confidence precision : perceptual precision ratios from functions fit **1)** to three datapoints (E.D.s with mean values set to the 2 confidence criterion values for that analysis, and to the perceptual category boundary – see Figure 5e), and **2)** to five datapoints (the 3 E.D.s

mentioned for the last analysis, and two more E.D.s with mean values set to +/- 1.5 x the two confidence criterion values). Note that both these sets of analyses are robust to violations of the normality assumption (due to E.D.s having excess kurtosis), as all confidence precision : perceptual precision ratios now fall within 5% of the target ratio of 1. So, these simulations suggest our Precision Test requires a minimum of 3 different test levels to be sampled (to target the two confidence criterion values and the perceptual category boundary – see Figure 5e).

**Discussion**

Simulations 1 and 2 have demonstrated that a popular SDT-based analysis of confidence (Maniscalco & Lau, 2012) can systematically underestimate metacognitive sensitivity when the normality assumption is violated, due to E.D.s being non-normally shaped either due to different skews (see Figure 3) or to excess kurtosis (see Figure 4). More importantly, these simulations also show that our new Precision Test to measure metacognitive sensitivity and confidence criteria is more robust to these violations of the normality assumption.

Simulation 3 showed that our new Precision Test is more robust to violations of the normality assumption because **1)** it does not implement a bespoke maximum likelihood estimation routine that is particularly susceptible to being undermined by violations of the normality assumption (see Figure 5a and 5c, and compare these to Figure 5b and 5d), and **2)** it samples more than 2 test values. Specifically, it samples test values that target important regions of an individual's psychological dimension when they make categorical perceptual decisions – that individual's two confidence criterion values and their perceptual category boundary (see Figure 2).

The Precision Test is robust to violations of the normality assumption, in large part because it aggregates data from across multiple test levels. The individual distributions that describe the different experiences people have of repeated presentations of each test level might individually have a non-normal shape, but once these data are aggregated from across ~5 test levels, a situation analogous to the central limits theorem (Fischer, 2010) seems to apply. Our precision test is consequently demonstrably more robust to violations of the normality assumption (see Figure 5) than are SDT-based analyses of confidence that do not aggregate data across this number of test levels (Maniscalco & Lau, 2012).

While we have established that our Precision Test is more robust to violations of the normality assumption than a currently popular SDT based analysis of confidence (Maniscalco & Lau, 2012), we need to validate the test procedure for use in experiments with humans. To this end, in Experiment 1 we re-examine an established relationship between the range of direction signals within a stimulus and perceptual confidence. When direction signals are generated by groups of moving dots, and people are asked to judge the average direction, confidence is disproportionately undermined (relative to the precision of perceptual decisions) when there is a large as opposed to a small range of different direction signals (de Gardelle & Mammassian, 2015; Spence et al., 2016). The results of these studies were unclear as to whether the root cause of this effect was a decline in metacognitive sensitivity when people judge more variable inputs, or if it was due to a confidence bias.

A subsequent study used a SDT-based analysis of confidence (Fleming, 2017), and found evidence that the impact of a large range of direction signals was to encourage a confidence

bias, with people more reticent to endorse categorical perceptual decisions regarding average direction with a relatively high level of confidence (see Spence et al., 2018). According to that evidence, participants should be equally able to distinguish between good and bad decisions regarding average direction when stimuli contain a broad or a restricted range of different direction signals (i.e. they should have equal levels of metacognitive sensitivity), but they should be biased to report having lower levels of confidence when inputs contain a large range of direction signals.

Given our concerns regarding SDT-based analyses of confidence, we thought it was worth re-examining the relationship between decisional confidence and the range of direction signals (de Gardelle & Mammassian, 2015; Spence et al., 2016) across two behavioural experiments with human participants. These experiments are matched on all details, except on how people report on confidence. In Experiment 1 people use a continuous scale to report on confidence, whereas in Experiment 2 they use a Likert scale to select one of a limited number of confidence ratings.

Another manipulation that can impact on measures of confidence is confirmation bias (e.g. Braun et al., 2018; Rollwage et al., 2020). When a preliminary perceptual decision is made, sampling of further perceptual evidence can become biased, with an amplification of decision congruent evidence, and insensitivity to decision incongruent evidence (Braun et al., 2018; Peters et al., 2017; Rollwage et al., 2020). To further validate our approach, in Experiment 3 we assess the impact of a confirmation bias on metacognitive sensitivity and bias.

## Experiment 1: Direction signal range, Metacognitive Sensitivity and Confidence Bias

This experiment was conducted to validate the use of the Precision Test of metacognitive sensitivity and bias in an experiment with human participants, and to re-assess the impact of the range of directional signals on metrics of metacognitive sensitivity and bias.

### Methods

A total of 26 volunteers participated, 16 female, with a mean age of 23 (S.D. 3.8). This would have delivered ~0.8 power to detect the effect size reported in the original study with a specified alpha of 0.01 (Spence et al., 2015). Data recordings failed for 3 participants, so the final sample consisted of 23 volunteer participants, with a mean age of 23 (S.D. 3.4). This should still deliver ~0.8 power to detect the effect size reported in the original study with a specified alpha of 0.05 (Spence et al., 2015). All participants were naïve as to the experimental hypotheses. Thirteen of the final group of volunteers participated in return for course credit. Ten were compensated with $40 for their participation. All participants completed experimental sessions while seated in a dimly lit room, viewing stimuli from a distance of 57 cm with their head restrained by a chin-rest. The study was approved by The University of Queensland research ethics committee, and was conducted in accordance with the principles of the Declaration of Helsinki.

### Stimuli

Stimuli consisted of random dot kinematograms (RDKs), generated using a Cambridge Research Systems ViSaGe stimulus generator driven by custom Matlab R2013b (MathWorks, Natick, MA) software and presented on a gamma-corrected 19 inch Dell P1130 monitor (resolution: 1600 x 1200 pixels; refresh rate: 60 Hz). Each RDK consisted of 101

individual white dots, each subtending 0.2 degrees of visual angle (dva) in diameter at the retinae, and drifting in a linear direction at a speed of 5 dva / second. Dots were presented against a black background within a circular aperture with a diameter subtending 4 dva. The individual lifetime of each dot was 200ms, after which it was re-drawn at a random position within the aperture. Each dot was assigned a random initial age (between 0 and 200ms).

In the center of the test display there was a small (diameter subtending 0.46 dva) red bull's eye configuration that served as the fixation point, which participants were instructed to stare at throughout all stimulus presentations.

Individual dots within RDKs translated in a uniform range of different directions (+/- 5 º or 20º) about a mean direction, which was slightly to the left or right of vertical. RDK presentations persisted for 1 second.

Immediately after stimulus presentations, participants were asked to simultaneously report on the average direction (left / right of vertical) and on their level of confidence in this decision (from a minimum rating of 5% on a linear scale, up to a maximum rating of 100%). These reports were made by moving a mouse left or right, by an amount that was indicative of the level of confidence. Mouse movements controlled the direction (left or right) in which a green bar was stretched, and how far the green bar was stretched (to indicate the level of confidence along a continuous scale). Participants would press a mouse button to finish their combined report. On alternate trials, RDKs moved predominantly up or down, and slightly to the left or right of vertical by a magnitude that was controlled via our adaptive sampling routines.

**Matlab scripts used to setup and control experimental sampling, and to analyze results**

Matlab scripts to setup, control and analyze the results of experiments that implement our Precision Test of metacognitive sensitivity and confidence bias are freely available as supplemental material and via our website Metacognitive Sensitivity: Precision Test (uq.edu.au). In Supplemental material we have also provided a document containing detailed descriptions of these scripts and their operation, with the aim that this can be used as a tutorial on the use and implementation of our scripts.

Our scripts set up and execute dynamic routines that target the sampling of test values to coincide with five regions of a psychometric function - the lower asymptote (where participants make high confidence decisions, that the average direction of inputs was *left* on ~90% of trials), the upper asymptote (where participants make high confidence decisions, that the average direction of inputs was *right* on ~90% of trials), the inflexion point (where participants are equally likely to report that the average direction had been left or right) and the mid-points in-between these three points.

In Experiment 1 there were two conditions, with tests either containing a uniform distribution of different direction signals ranging +/-5 º or +/-20º from the average test direction. During a single experimental session, completed by each participant, 200 individual trials were conducted for each condition. At the conclusion of an experimental session, our routines fit three psychometric functions to data for each condition – as depicted in Figure 2, and described in the subsection titled ‘A new Precision Test of metacognitive sensitivity and confidence bias’. From these, a high confidence criterion value is estimated from the average horizontal separation of the two psychometric functions that measure confidence (see the blue

and red functions depicted in Figure 2) from a central function that describes the proportion of trials on which the average test direction had been categorised as moving to the right of vertical (regardless of confidence, see the black functions Figure 2).

At the conclusion of an experimental session, for each experimental condition our code calculates the ratio of the average of the two standard deviations of functions fit to data describing confidence (see the blue and red functions depicted in Figure 2) relative to the standard deviation of the function fit to data describing perceptual decisions (regardless of confidence, see black functions in Figure 2). If a participant is metacognitively ideal, this ratio should equal 1. If a participant is metacognitively insensitive, because confidence ratings have been shaped by additional information or noise relative to perceptual decisions, the precision ratio should be equal to some value less than 1.

**Results**

The results of Experiment 1 are depicted in Figure 6. Prior to calculating statistics and plotting data, each dataset was analysed to detect outlier values (> +/- 3 S.D.s from the dataset mean), and these were removed (so outlier values are not plotted, and they have no impact on statistical analyses). This is true of all the behavioural datasets we report. Code used to analyse the results of our behavioural experiments (1 – 3) and to generate Figures 6 – 8 is provided as supplemental material.

Metacognitive precision ratios were less than 1 for both of our experimental conditions (see Figure 6a). This suggests that confidence ratings were shaped by different information or noise, relative to perceptual category decisions. There was, however, no evidence for a robust

difference between the two conditions, with a Bayes factor analysis revealing moderate evidence for the null hypothesis, that there would be no difference in metacognitive precision ratios across the two experimental conditions (see Figure 6a).

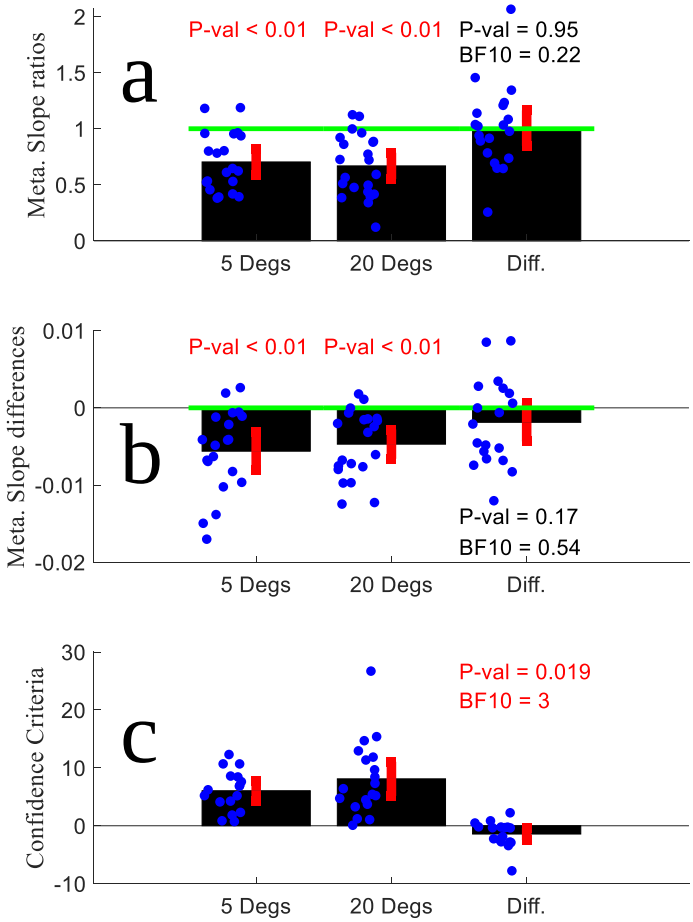


**Figure 6.** Results of Experiment 1. **a)** Bar plots, depicting Confidence : Perceptual function fit precision ratios, for RDKs with a uniform range of direction signals +/- 5° from the mean (5 Degs) or +/- 20° from the mean (20 degs), along with values of 1 - the difference between these two ratios (Diff). Blue data points depict individual ratios, and red vertical bars depict +/- 2SEM from the group average. The green horizontal line marks a ratio of 1, which is consistent with metacognitively ideal performance (for the 5 and 20 Degs datasets), and with there being no difference in metacognitive sensitivity across these two conditions (Diff). **b)** Details are as for Figure 6a, but for data relating to precision differences. Here, differences of 0 are consistent with metacognitively ideal performance, and with there being no differences across the two conditions. **c)** Details are as for Figure 6b, but for data relating to confidence criteria. These data do not speak to metacognitive sensitivity.

We have also expressed metacognitive precision relationships as conditional difference scores (Avg. Confidence Precision - Perceptual Precision, see Figure 6b). Average

differences for both experimental conditions were < 0, suggesting that confidence ratings were shaped by different information or noise relative to perceptual category decisions. There was, however, again no evidence for a robust difference between metacognitive precision differences across the two experimental conditions, with a Bayes factor analysis revealing moderate evidence for the null hypothesis, that there would be no conditional difference (see Figure 6b).

The results of Confidence criterion analyses are depicted in Figure 6c. People adopted a more conservative confidence criterion (they needed a greater average direction offset from vertical before they would report having a relatively high level of confidence) for RDKs that contained a broader range of directions signals (see Figure 6c). A Bayes factor analysis revealed a moderate level of evidence for the alternative hypothesis, that people would adopt different confidence criteria when judging the average direction of RDKs that have different ranges of direction signal. Overall, these data suggest the impact on confidence of having a larger range of direction signals is to cause people to be more conservative when deciding if they should endorse a perceptual decision with a relatively high-level of confidence. These data suggest there is no adverse impact on metacognitive sensitivity into decisions regarding average direction.

In Experiment 1 people made simultaneous direction decisions and confidence ratings, by choosing a direction in which to stretch a bar (left or right) to indicate the categorical direction decision, and by choosing how far to stretch that bar to indicate the level of confidence (along a continuous scale) they felt in relation to this decision. Many studies use

sequential reports instead, and a Likert scale to report on confidence. In Experiment 2 we wanted to determine if the Precision Test would be robust to these changes.

## Experiment 2: Direction signal range, Metacognitive Sensitivity and Confidence Bias measured using a Likert scale and sequential decisions

This experiment was conducted to further validate the use of our Precision Test with human participants. Importantly, we show that the Precision Test can be used in Experiments where people report on confidence using a Likert scale, as opposed to a continuous measure of confidence.

All details for Experiment 2 were as for Experiment 1, with the following exceptions. Immediately after test presentations people were asked to press one of two buttons to indicate if the mean test direction had been to the left or right of vertical. After this response, they were asked to rate their level of confidence by using a mouse to highlight and click on one of 5 confidence ratings labelled ‘1 (Guessing’, ‘2’, ‘3’, ‘4’, or ‘5 (Certain)’.

### Results

As in Experiment 1, metacognitive precision ratios were less than 1 for both experimental conditions (see Figure 7a), suggesting confidence ratings were shaped by different information or noise, relative to perceptual category decisions. Again, there was also no evidence for a robust difference between conditions in terms of metacognitive sensitivity, with a Bayes factor analysis revealing moderate evidence for the null hypothesis, that there would be no conditional difference in metacognitive precision ratios (see Figure 7a).

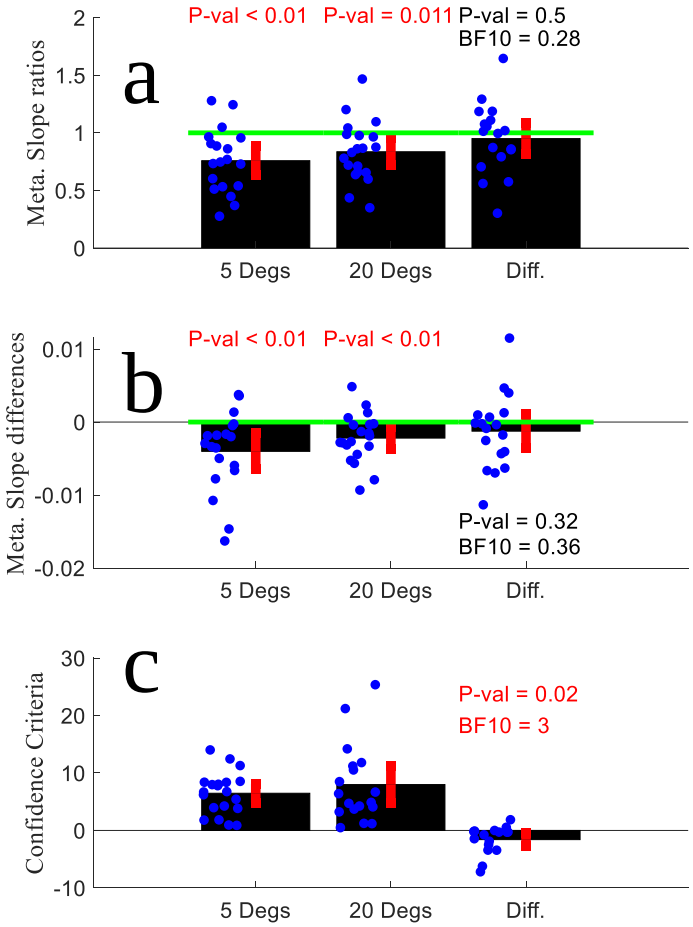


**Figure 7.** Results of Experiment 2. All details regarding this figure are as for Figure 6.

As per Experiment 1, we also expressed metacognitive precision relationships as conditional difference scores (see Figure 7b). Both average conditional difference scores were < 0, suggesting confidence ratings were shaped by different information or noise relative to perceptual direction decisions. There was, however, no robust evidence for there being a difference between these conditional differences (see Figure 7b).

The results of Confidence criterion analyses are depicted in Figure 7c. Again, as in Experiment 1, people adopted a more conservative confidence criterion for tests that contained a broader range of direction signals (see Figure 7c), with a Bayes factor analysis revealing a moderate level of evidence for the alternative hypothesis – that people would

adopt different confidence criteria when judging the average direction of tests that contain different ranges of direction signals.

The results of Experiment 2 reiterate those of Experiment 1, reinforcing the view that the impact of a larger range of direction signals on confidence is to cause people to be more conservative when deciding if they should endorse a perceptual category decision with a relatively high-level of confidence.

The results of Experiments 1 and 2 conceptually replicate the results of an earlier study, which had suggested a larger range of direction signals encourages people to be more conservative when deciding if they should endorse a perceptual direction decision with a relatively high level of confidence (Spence et al., 2018). In that study, and in Experiments 1 and 2 here, there was no robust evidence for a detrimental impact on metacognitive sensitivity.

In Experiment 3, we attempt to validate our Precision Test in a context where other studies have suggested a conditional difference in metacognitive sensitivity should exist. We sought to encourage selective attention to directional information, by having people guess at what the average direction would be in a subsequent test *before* it was presented.

## Experiment 3: Direction Perception, Confidence and Confirmation Bias

This experiment was conducted to further validate the Precision Test. We wanted to use it to assess metacognitive sensitivity and confidence criteria in a context where evidence has suggested there should be a conditional difference in metacognitive sensitivity.

Details for Experiment 3 were as for Experiment 1, with the following exceptions. All tests contained a range of direction signals +/- 30º from the test mean, and prior to test presentations people were asked to *guess* what the average test direction would be. They were encouraged to consider their last 3 guesses (Left or Right) in order to 'help' them make a prediction. A record of these decisions was displayed, initialized with place-holding 'none' labels that were progressively filled as trials were completed and predictions made. Our hope was to elicit a gambler's fallacy-like (e.g. Xu & Harvey, 2014) sense that the participant could anticipate these randomized stimulus presentations, by considering the pattern of recent presentations. Experimental conditions refer to tests that were Consistent with guesses, or Inconsistent with guesses.

### Results

The results of Experiment 3 are depicted in Figure 8. Metacognitive precision ratios were not, on average, less than 1 (see Figure 8a). Nor was there any robust evidence for a difference between conditions in terms of metacognitive sensitivity (see Figure 8a). We have also expressed metacognitive precision relationships as conditional difference scores (see Figure 8b). In this analysis there was evidence that Confidence functions were *less precise* than Perceptual functions when tests were consistent with guesses, whereas there was no evidence for this difference when tests were inconsistent with guesses. There was also some evidence

for a conditional difference between these two sets of conditional difference scores, with evidence for reduced precision when tests were consistent with guesses.

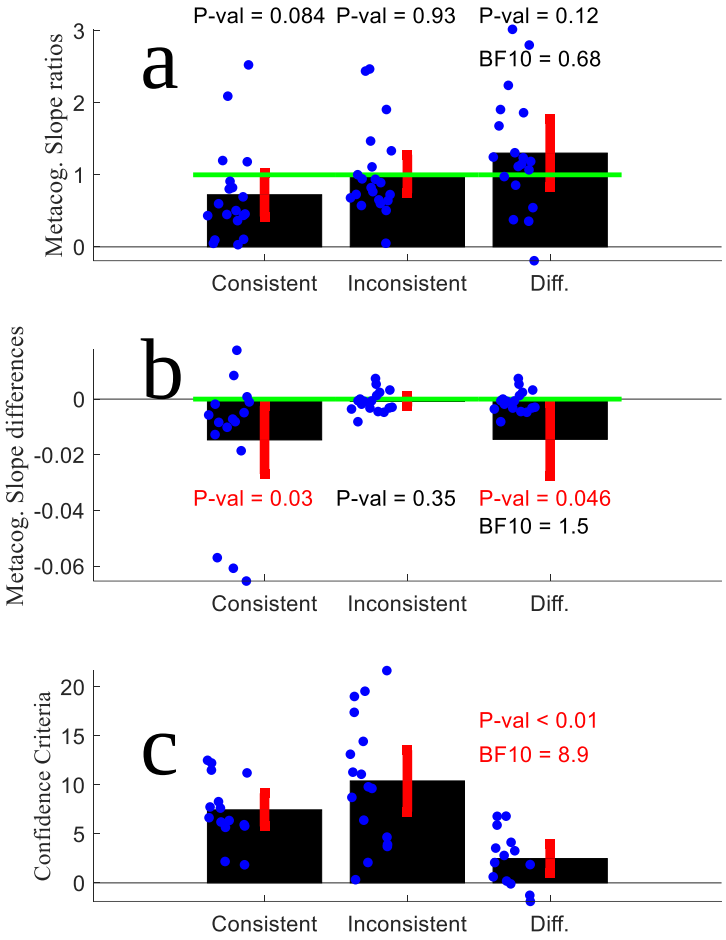


**Figure 8.** Results of Experiment 3. All details regarding this figure are as for Figure 7, except that experimental conditions refer to when tests that were Consistent with guesses concerning what the average test direction would be, or to when tests were Inconsistent with guesses concerning what the average test direction would be.

The results of Confidence criterion analyses are depicted in Figure 8c. These data show that people adopted a more conservative confidence criterion when tests were Inconsistent with their guesses (as to what the average test direction would be, see Figure 8c). A Bayes factor analysis revealed a moderate level of evidence for the alternative hypothesis, that people would adopt different confidence criteria when judging the average direction of tests that were Consistent or Inconsistent with their guesses.

## General Discussion

We have described a new Precision Test to estimate metacognitive sensitivity and confidence criteria. In simulations we showed that this method (see Figures 3-5) is more robust to violations of the normality assumption than an established SDT-based analysis of confidence (Maniscalco & Lau, 2012). We then validated the new method in a series of behavioural experiments with human participants. Data from these experiments suggest that having a broader range of direction signals causes people to be more conservative when they rate their confidence in direction decisions (a confidence bias effect, see Figures 6-7). Metacognitive sensitivity (their ability to distinguish between likely correct and inaccurate decisions) is undiminished (see Figures 6-7, also see Spence et al., 2018). In Experiment 3, looking at the impact of conformation bias, we found evidence that this can both selectively reduce metacognitive sensitivity for decisions regarding expected inputs (see Figure 8b), and cause people to be more conservative when rating their confidence in decisions about expected inputs (see Figure 8c).

### *Differences between the Precision Test and fitting separate psychometric functions to data associated with Low and with High Confidence decisions.*

A number of studies have fit separate psychometric functions to data associated with relatively low and with relatively high levels of decisional confidence (e.g. de Gardelle & Mamassian, 2014; De Martino et al., 2013; Keane et al., 2015; Mamassian & de Gardelle, 2022). As these studies aggregate information from across multiple test levels when they model data, so they should also benefit from a situation analogous to the central limits theorem (Fischer, 2010) and be more robust to violations of the normality assumption than

are analyses that are informed by just two test levels. However, this type of analysis is quite different to our Precision Test, and it does not easily deliver the same levels of insight.

The confidence functions of our Precision test pivot about confidence criteria, and so they deliver an estimate of a function that describes transitions between endorsing decisions with a relatively low to a relatively high level of confidence. As these functions do not pivot about and directly reference the perceptual category boundary, they will deliver estimates of the precision of confidence judgments that are somewhat independent of the precision of perceptual category decisions. Functions fit to datasets that are merely split by confidence (e.g. de Gardelle & Mamassian, 2014; De Martino et al., 2013; Keane et al., 2015; Mamassian & de Gardelle, 2022) still pivot about the perceptual category boundary, so they confound perceptual category and confidence noise.

The data that inform 'high' confidence functions in data split analyses were not necessarily associated with a high level of confidence. Particularly when people are forced to choose which of two sequential decisions had been associated with most confidence (e.g. de Gardelle & Mamassian, 2014; Mamassian & de Gardelle, 2022) low confidence decisions will often inform the nominally 'high' confidence dataset (and high confidence decisions will inform nominally 'low' confidence dataset). These decisions need only elicit more confidence than the comparison judgment on a pair of trials, or they may be chosen at random when decisional confidence levels are indistinguishable. In our Precision Test, participants are allowed to express an intrinsic level of felt confidence, and so we know that high confidence decisions were associated with a relatively high level of felt confidence.

The horizontal separation of the two confidence functions in our Precision Test also delivers an estimate of the participants' confidence criterion – an estimate of how big a categorical difference must be before a participant will endorse that perceptual category decision with a relatively high level of confidence. This is not delivered by fitting separate psychometric functions to low and to high confidence datasets.

*Differences between the Precision Test and Type-2 AROC calculations.*

A long standing and popular means of estimating metacognitive insight is to evaluate the proportions of correct and incorrect decisions when people express different levels of confidence in their decisions (Benjamin & Diaz, 2008; Galvin et al., 2003). The different levels of confidence can be treated as different confidence criteria to repeatedly split data (between decisions associated with low and with high confidence, relative to the nominal confidence criterion). For each confidence split, proportions of correct (i.e. the proportion of correctly classified 'Signal' presentations) and incorrect (wrongly classified 'Noise' presentations) 'high' confidence decisions can be evaluated and plotted as cartesian datapoints. The area under a curve fitted to the plotted datapoints can then be taken as a metric of metacognitive sensitivity – as it can indicate the degree to which relatively high confidence decisions were more often correct. This procedure is referred to as a Type-2 Area under a Receiver Operating Curve *(Type-2 AROC)* analysis (see Fleming & Lau, 2014). This approach does not commit to any assumptions regarding the shape the experiential distributions that underlie these analyses, so it is robust to violations of the normality assumption. However, this approach is undermined by other issues that do not impact on our Precision Test.

First, metacognitive sensitivity estimates derived from area under the curve calculations can be relatively insensitive, as a reasonably large change in metacognitive sensitivity can have only a slight impact on area calculations. Second, confidence contingent area under the curve calculations do not deliver a metric of metacognitive sensitivity that is directly comparable to metrics of the underlying task performance. Our Precision Test, however, delivers directly comparable metrics – the standard deviations that describe confidence and perceptual task performance (see Figure 2).

*Researchers will likely continue to use SDT-based analyses of confidence – for good reasons!*

While we believe our Precision Test of metacognitive sensitivity is superior, in terms of being more robust to violations of the normality assumption and less susceptible to interpretive ambiguities (Arnold et al., 2023; Miyoshi et al., 2022), we nonetheless expect researchers will continue to use SDT-based experimental designs and analyses of confidence (e.g. Fleming, 2017; Fleming and Lau, 2014; Maniscalco & Lau, 2012, 2016). One good reason is that a large number of repeated presentations of a small number of inputs can be needed to support fMRI and EEG analyses. This necessity can preclude the sampling of a greater number of different test levels, which is needed to implement our Precision Test. In some circumstances, the cost of an interpretive ambiguity might by outweighed by the practical need to sample a large number of repeated presentations of inputs.

*Things to be aware of if you use our method*

Our method is easy to implement in standard experimental designs. We have also provided code to make this implementation easy for other researchers: see supplemental material, and

Metacognitive Sensitivity: Precision Test (uq.edu.au). We hope this will be useful to others. There are, however, issues you should be aware of.

Our approach requires a sampling of inputs that result in people reliably making High-Confidence Category A responses at one extreme of the test range, and High-Confidence Category B responses at the other extreme, with intervening test values resulting in variable Low-Confidence category decisions. If there is no variance in a participants' confidence ratings, our approach cannot be implemented.

We would suggest it is worth conducting a little training to screen participants prior to formal data collection (this process and exclusion criteria should be pre-registered). A participant could be shown exemplar inputs that should be trivial to categorise (say stimuli that are tilted +/- 30 degrees from vertical, or some equivalent for your experimental design). The experimenter would need to confirm that each participant can reliably categorise these exemplars, and if they can, instruct the participant that if these inputs can be clearly discerned, they should report having a relatively high-level of confidence in these decisions. Participants could also be informed that other inputs will be ambiguous, and that these should evoke a confidence rating that corresponds with the ambiguity of the input.

Without such training we have found that many participants will never reliably endorse perceptual category decisions with high confidence. Perhaps they have failed to understand task instructions, or they think there is some unknown trick to the experiment that undermines their confidence. Regardless of the reason, if people will not reliably endorse a blatantly obvious perceptual category decision with a relatively high level of confidence, our

approach will fail. But perhaps in a good way. Our data plots will reveal when a participant has failed to understand task instructions, whereas this can be less obvious in other experimental approaches.

Like other approaches to estimating metacognitive sensitivity (e.g. Fleming, 2017; Fleming and Lau, 2014; Maniscalco & Lau, 2012, 2016), our Precision Test allows researchers to express the precision of confidence decisions as either a ratio (relative to the precision of perceptual category decisions), or as a difference score. Given that ratios can multiply measurement errors, researchers would be well-advised to use difference scores. Indeed, the results of Experiment 3 show that difference scores can provide a more robust basis for analyses.

*What do our data say about confidence in direction decisions?*

Our experiments were primarily conducted to validate our new Precision Test of metacognitive sensitivity and confidence criteria, through behavioural experiments with human participants. Our data do, however, allow us to make some observations.

First, we have found evidence that a wider range of direction signals impacts on confidence by encouraging participants to be more conservative when deciding if they should endorse a perceptual category decision with a relatively high level of confidence (see Figures 6-7). We have found no evidence for an impact on metacognitive sensitivity. These findings replicate those of an earlier study that used a different measure of the impact of confidence (Fleming, 2017) but reached the same conclusions (Spence et al., 2018). We also found evidence that an encouraged confirmation bias, by having people guess at what the average direction will be

within a future test, can reduce metacognitive sensitivity for bias-confirming evidence, and induce a confidence bias for confirmatory evidence (as per Braun et al., 2018; Peters et al., 2017; Rollwage et al., 2020). In each case our data have provided convergent evidence in relation to existing findings.

*Conclusions*

We have described a novel means of measuring metacognitive sensitivity and confidence criteria. This Precision Test for metacognitive sensitivity and confidence criteria is more robust against violations of the normality assumption than the currently most popular SDT-based analysis of confidence (Maniscalco & Lau, 2012). We have provided code researchers can access to implement our procedures as supplemental material, and at a lab website https://www.psy.uq.edu.au/~uqdarnol/MetaCog_Precision_Test.html.

**References**


Acerbi L., Wolpert D.M. & Vijayakumar S. (2012). Internal representations of temporal statistics and feedback calibrate motor-sensory interval timing. *PLoS Computational Biology* **8(11),** e1002771

Anderson B. (2014). Increased kurtosis for judgements of probable feature/position conjunctions. *Journal of Vision* **15,** 1 – 11.

Appelle S. (1972). Perception and discrimination as a function of stimulus orientation: The "oblique effect" in man and animals. *Psychological Bulletin*, **78,** 266 – 278.

Arnold, D.H., Johnston, A., Adie, J. & Yarrow, K. (2023). On why we lack confidence in some signal-detection-based analyses of confidence. *Consciousness & Cognition*, **113,** 103532.

Arnold, D.H. Saurels, B.W., Andersen, N.L. & Johnston, A. (2021). An observer model of tilt perception, sensitivity and confidence. *Proceedings of the Royal Society of London B* **288,** 1 – 8.

Barlow, H. B. (1962). A method of determining the over-all quantum efficiency of visual discriminations. *The Journal of Physiology*, **160,** 155–168.

Bays P.M. (2016). A signature of neural coding at human perceptual limits. *Journal of Vision* **16(11),** 4, 1 – 12.

Braun, A., Urai, A. E., & Donner, T. H. (2018). Adaptive history biases result from confidence-weighted accumulation of past choices. *Journal of Neuroscience* **38,** 2418 – 2429.

Clifford C.W.G., Wenderoth P. & Spehar B. (2000). A functional angle on some after-effects in cortical vision. *Proceedings of the Royal Society of London B* **267,** 1705 – 1710.

de Gardelle, V. & Mamassian, P. (2014). Does confidence use a common currency across two visual tasks? *Psychological Science*, **25(6)**, 1286 – 1288.

de Gardelle, V. & Mamassian, P. (2015). Weighting mean and variability during confidence judgments. *PLoS ONE*, **10(3)**, e0120870

De Martino, B., Fleming, S.M., Garrett, N. & Dolan, R.J. (2013). Confidence in value-based choice. *Nature Neuroscience*, **16(1)**, 105 – 110.

Fischer H. (2010). A history of the central limit theorem: From classical to modern probability theory. New York: Springer.

Fleming, S. M. (2017). HMeta-d: Hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness 2017*, **nix007**.

Fleming S.M., Dolan R.J. & Frith C.D. (2012). Metacognition: computation, biology and function. *Philosophical Transactions of the Royal Society of London B* **367,** 1280-1286.

Fleming, S.M. & Lau, H.C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience* **8,** 443.

Geisler, W.S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological Review* **96,** 267 – 314.

Gibson, J. J. & Radner, M. J. (1937). Adaptation, after-effect, and contrast in the perception of tilted lines. Quantitative studies. *Journal of Experimental Psychology* **20,** 453–467.

Girshick A.R., Landy M.S. & Simoncelli E.P. (2011). Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience* **14,** 926–932.

Green, D.M. & Swets, J.A. Signal detection theory and psychophysics. Wiley, New York (1966).

Jabar S.B. & Anderson B. (2015). Probability shapes perceptual precision: A study in orientation estimation. *Journal of Experimental Psychology: Human Perception and Performance* **41,** 1666 – 1679.

Keane B., Spence M., Yarrow K. & Arnold D.H. (2015). Perceptual confidence demonstrates trial-by-trial insight into the precision of audio–visual timing encoding. *Consciousness and Cognition* **38,** 107 – 117.

Levitt, H.L. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America* **49,** 467 – 477.

Li Q., Hill Z. & He B.J. (2014). Spatiotemporal dissociation of brain activity underlying subjective awareness, objective performance and confidence. *The Journal of Neuroscience* **34,** 4382 – 4395.

Mamassian, P. & de Gardelle, V. (2022). Modeling perceptual confidence and the confidence forced-choice paradigm. *Psychological Review* **129,** 976 – 998.

Maniscalco B. & Lau H.C. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition* **21,** 422 – 430.

Maniscalco B. & Lau H. (2016). The signal processing architecture underlying subjective reports of sensory awareness. *Neuroscience of Consciousness, 2016,* **Article niw002**.

Maniscalco, B., Peters, M. A., & Lau, H. (2016). Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Attention, Perception, & Psychophysics* **78,** 923-937.

Masson, M. E., & Rotello, C. M. (2009). Sources of bias in the Goodman–Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **35,** 509.

Mazorm M., Friston, K.J. & Fleming, S.M. (2020). Distinct neural contributions to metacognition for detecting, but not discriminating visual stimuli *eLife* **9,** e53900.

Miyoshi K., Sakamoto Y. & Nishida S. (2022). On the assumptions behind metacognitive measurements: Implications for theory and practice. *Journal of Vision* **22,** 18, 1 – 15.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin* **95,** 109 – 133.

Peters M.A.K., Thesen T., Ko Y.D., Maniscalco B., Carlson C., Davidson M., Doyle W., Kuzniecky R., Devinsky O., Halgren E., & Lau H. (2017). Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature Human Behavior* **1,** 7, 1 - 8.

Rausch, M., Müller, H. J., & Zehetleitner, M. (2015). Metacognitive sensitivity of subjective reports of decisional confidence and visual experience. *Consciousness and cognition*, **35,** 192-205.

Regan D. & Beverley K.I. (1985). Postadaptation orientation discrimination. *Journal of the Optical Society of America A* **2,** 147–155.

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,** 2634.

Roseboom, W., Fujisaki, W., Nishida, S. & Arnold, D.H. (2011). Audio-visual speech timing sensitivity is enhanced in cluttered conditions. *PLoS One* **6(4),** e18309.

Samaha, J. & Denison, R. (2022). The positive evidence bias in perceptual confidence is unlikely post-decisional. *Neuroscience of Consciousness* **2022,** 1.

Spence, M. L., Mattingley, J. B., & Dux, P. E. (2018). Uncertainty information that is irrelevant for report impacts confidence judgments. *Journal of Experimental Psychology: Human Perception and Performance* **44,** 1981–1994.

Spence, M., Dux, P. & Arnold, D.H. (2016). Computations Underlying Confidence in Visual Perception. *Journal of Experimental Psychology: Human Perception & Performance* **42,** 671 – 682.

Storrs K.R. & Arnold D.H. (2015a). Face after-effects involve local repulsion, not renormalization. *Journal of Vision* **15(8),** 1, 1 – 18.

Storrs K.R. & Arnold D.H. (2015b). Evidence for tilt normalization can be explained by anisotropic orientation sensitivity. *Journal of Vision* **15(26),** 1, 1 – 11.

Webster, M.A. (2015). Visual Adaptation. *Annual review of vision science. U.S.A.* **1,** 547-567.

Xu, J. & Harvey, N. (2014). Carry on winning: The gamblers' fallacy creates hot hand effects in online gambling. *Cognition* **131,** 173-180.

Yarrow, K., Jahn, N., Durant, S. & Arnold, D. H. (2011). Shifts of criteria or neural timing? The assumptions underlying timing perception studies. *Consciousness & Cognition* **20,** 1518–1531.

Yeung N. & Summerfield C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society of London B* **367,** 1310-1321.