On why we lack confidence in some signal-detection-based analyses of confidence

Derek H. Arnold[1], Alan Johnston[2], Joshua Adie[3] & Kielan Yarrow[4]

[1]School of Psychology, The University of Queensland; [2]School of Psychology, The University of Nottingham; [3] Research Institute for Sport & Exercise, University of Canberra; [4]Department of Psychology, City University London

Signal-detection theory (SDT) is one of the most popular frameworks for analyzing data from studies of human behavior – including investigations of confidence. SDT-based analyses of confidence deliver both standard estimates of sensitivity (d'), and a second estimate informed by high-confidence decisions – meta d'. The extent to which meta d' estimates fall short of d' estimates is regarded as a measure of metacognitive inefficiency. These analyses rely on a key but questionable assumption – that repeated exposures to an input will evoke a normally-shaped distribution of perceptual experiences (the normality assumption). We show that when distributions of experience do not conform with the normality assumption, meta d' can be systematically underestimated relative to d'. We explain why violations of the normality assumption are especially a problem for some popular SDT-based analyses of confidence, in contrast to other analyses inspired by the SDT framework, which are more robust.

Key Words: Perceptual metacognition; Confidence; Visual Adaptation; 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 show that the results of these analyses cannot be regarded as ground truth. If a key assumption of the framework is inadvertently violated, analyses can encourage conceptually flawed conclusions.

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. & A.J.

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

Author contributions: D.H.A & A.J. conceived of the study. D.H.A. programmed experiments, analysed data and wrote the first draft of the manuscript. J.A. tested participants. All authors edited successive versions of the manuscript.

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

Introduction

When people make decisions, we experience subjective levels of confidence regarding the quality of our decisions (Fleming et al., 2012; Yeung & Summerfield, 2012). In perception, these feelings are typically accurate, with higher levels of confidence associated with more accurate decisions (Keane et al., 2015; Li et al., 2014; Peters et al., 2017). This capacity, to report on the accuracy of our intrinsic decisional processes, is known as perceptual metacognition.

Historically, quantifying the degree of insight we might have into the internal operations of our minds has been challenging. The recent extrapolation of a standard framework, traditionally used to estimate perceptual sensitivity (Green & Swets, 1966), to appraise confidence has therefore been noteworthy – promising an objective measure of the degree of insight we might have into the quality of our perceptual decisions (Fleming & Lau, 2014; Maniscalco & Lau, 2012).

To understand the new approach, we need to start with a reprisal of standard implementations of the SDT framework (Green & Swets, 1966). Popular implementations of the framework commit to a set of assumptions that allow investigators to calculate independent estimates of objective sensitivity (d'), and of the subjective criteria people use when making perceptual decisions. One of the key assumptions is that repeated exposures to a given physical input will be associated with a normally-shaped distribution of different perceptual experiences (Barrett et al., 2013; Fleming & Lau, 2014; Green & Swets, 1966; Maniscalco & Lau, 2012; Yarrow et al., 2011). For instance, a vertical input might most often be perceived as vertical, but also be seen as differently tilted, left and right of vertical on some trials. It is assumed that the various experiences following exposures to a common physical input will be normally distributed. It is

this assumption that allows researchers to make a backwards inference, from behavioral data, to estimate the areal overlap between theoretical distributions that relate to the different experiences that people have had following exposures to a target (signal) and a non-target (noise) input (see Figure 1).

To assess confidence, the standard SDT framework has been extended to include additional 'confidence' criteria – values against which experiences are referenced to assign levels of confidence (see Figure 1b). In a minimal case, you need two additional confidence criteria, to demark low- from high-confidence decisions. You could, for instance, have a criterion to demark low- from high-confidence target categorisations (see the bold black vertical bar, Figure 1b). With the addition of these criteria, researchers can backwards infer a second estimate of how separated non-target noise and target signal experiential distributions must be, from analyses informed by proportions of correct (e.g. Figure 1b, light green shaded area) and incorrect (e.g. Figure 1b, dark green shaded area) high-confidence decisions (Fleming & Lau, 2014; Maniscalco & Lau, 2012). In essence, this process estimates the same property as more traditional SDT-based analyses (d'), but from an estimate of the overlap of the *tails* of target and non-target distributions (see Maniscalco & Lau, 2012). The key statistic this process delivers is meta d'.

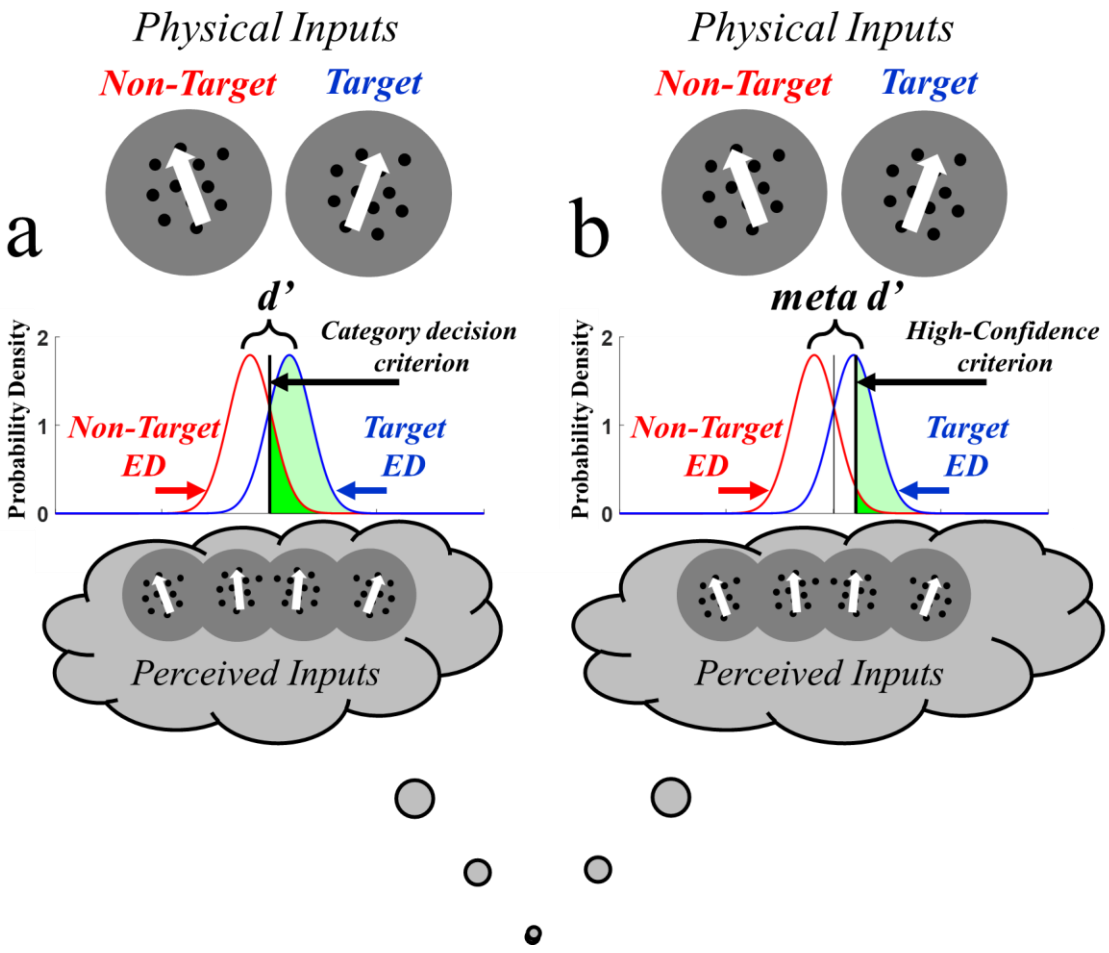

Figure 1. a) Graphic depicting popular implementations of the SDT framework. Repeated presentations of a non-target and of a target input (here dots moving left and right of vertical – see top of graphic) will result in different perceptual experiences (different perceived directions – see bottom of graphic). These are thought to be described by normally-shaped distributions – here labelled as Experiential Distributions (EDs). It is further assumed that people reference experiences against a category criterion value (vertical black bar) to demark non-target (values to the left of the criterion) from target experiences (values to the right of the criterion). These assumptions allow researchers to backwards infer how separated target and non-target EDs must be in order to be consistent with the proportion of non-target presentations categorized as targets (the false alarm rate – see dark green shaded region) and with the proportion of target presentations categorized as targets (the hit rate – see light green shaded region). This estimate, d', is regarded as a measure of perceptual sensitivity. **b)** Graphic depicting assumptions underlying popular extensions of the SDT framework, used to analyze confidence. A high-confidence criterion (bold black vertical line) is assumed to demark experiences classified as having been of targets with low confidence (values immediately to the left of this criterion) or with high confidence (all values to the *right* of the high-confidence criterion). This criterion is applied to both non-target (dark green shaded region) and target (light green shaded region) presentations (a mirror criterion is assumed to demark experiences classified as non-targets with low or high confidence, but this is not illustrated). By assuming the existence of confidence criteria, researchers can backwards infer how separated target and non-target EDs must be, in order to be consistent with proportions of *correct* (e.g. dark green shaded region) relative to *incorrect* (e.g. light green shaded region) high-confidence categorizations. This estimate is known as meta-d'.

When meta d' and d' estimates are equivalent, an observer is said to be metacognitively 'ideal', in that the data are consistent with d' and meta d' estimates having been informed by a common source of information. The degree to which an observer is said to be metacognitively inefficient is thought to be indicated by the degree to which meta d' estimates fall short of d' estimates (Fleming & Lau, 2014; Maniscalco & Lau, 2012). This situation is typically regarded as evidence that estimates of confidence have been contaminated by an additional source of noise, relative to perceptual judgments (Fleming & Lau, 2014; Maniscalco & Lau, 2012). Standard implementations of this framework rest on the assumption that distributions of experiences after repeated exposures to an input are normally-shaped, such that areal overlaps of distributions can be inferred from behaviour.

A structural concern for SDT-based analyses of confidence is that we don't know the precise shape of experiential distributions, and there is good reason in visual perception to suspect these might often deviate from normality. The assumption of normally-shaped experiential distributions presumes that an input is equally likely to be mis-perceived to either side of its veridical position along a psychological dimension. A near vertical input, for instance, should be equally likely to be misperceived as vertical or as further tilted from vertical. Evidence, however, suggests systematic biases, with some studies suggesting that near vertical inputs are more likely to be misperceived as vertical than as further titled from vertical (Appelle, 1972; Bouma & Andriessen, 1968; Girshick et al., 2011), while other studies have suggested the opposite bias (e.g. Tomassini et al., 2010). In either case, this would involve localised skews within the psychological dimension (Appelle, 1972; Girshick et al., 2011; Storrs & Arnold., 2015b).

In addition to an absence of skew, the assumption of normally-shaped experiential distributions (Fleming & Lau, 2014; Maniscalco & Lau, 2012) presumes that experiential distributions should be mesokurtic, such that the shapes that describe the tails of an experiential distribution should be matched to those that describe the tails of a normal distribution. However, it has been established that detailed datasets describing human decisions can be equally well explained by decisions having been informed by a number of differently shaped experiential distributions (see Rouder et al., 2010). This observation has been advanced as motivation to avoid excessive dependence on assuming a particular shape of experiential distribution within analyses of perceptual decisions (e.g. Kellen & Klauer, 2015; Miyoshi et al., 2022).

Even if inputs were usually mapped onto different perceptual experiences, mappings between inputs and perception can be made to undergo temporary changes, like those invoked by visual adaptation (by prolonged exposure to a given physical input – such as upwards motion). The most obvious changes following visual adaptation are to perception – with a tendency to see similar inputs as repelled from the adaptor (Clifford et al., 2000; Gibson & Radner, 1937; Regan & Beverly, 1985; Webster, 2015). Adaptation to upwards motion, for instance, can make movements to the left and right of vertical seem to be moving in directions more rotated from upwards (Clifford et al., 2000; Webster, 2015). These perceptual changes are accompanied by less obvious changes to the precision of perceptual judgments, with the direction and magnitudes of both types of change being governed by the proximity of inputs to the adaptor within the psychological dimension (Clifford et al., 2001; Regan & Beverly, 1985; Webster, 2015). Recently, it has additionally been shown that visual adaptation can impact measures of perceptual confidence (Arnold et al., 2021).

A benefit of visual adaptation is that we can leverage knowledge of the processes underlying human vision to produce biologically-inspired population coding models, that detail how changes in perception, in the precision of perceptual decisions, and in confidence, might all have been produced (Clifford et al., 2001; Jin et al., 2005; Kohn & Movshon, 2004; Kohn, 2007; Storrs & Arnold, 2012; Storrs & Arnold, 2015a,b). An important conceptual point is that the success of these models depends on a description of how adaptation might have temporarily changed mappings between inputs and perception, with distinct changes prevailing at different loci along a psychological dimension. Our recent study showed that this approach could describe how tilt adaptation impacts measures of human perception, perceptual precision, and most importantly confidence – specifically the spread of uncertainty. Uncertainty tended to be increased by adaptation (i.e. a greater range of test tilt differences resulted in low confidence decisions), and this impact was greater in magnitude than the accompanying adverse changes to the precision of tilt decisions (Arnold et al., 2021).

An additional advantage of biologically inspired population coding models of perception (Clifford et al., 2001; Jin et al., 2005; Kohn & Movshon, 2004; Kohn, 2007; Storrs & Arnold, 2012; Storrs & Arnold, 2015a,b), untapped until now, is that these can be quizzed to explore what shape experiential distributions might have, before and after visual adaptation. We can repeatedly input orientations into a model (in either an unadapted baseline state, or an adapted state), and see what shapes describe the distributions of encoded orientations (which are hoped to mimic human perception), and more importantly, we can test if any deviance from the normality assumption is sufficient to undermine SDT-based analyses of confidence.

Here we will report on a study that had four stages. First, we demonstrate the robustness of behavioural results from our past investigation, by assessing the impact of visual adaptation on confidence in a different perceptual dimension (direction perception). To preface these results, confidence is again robustly impacted by adaptation. In a second stage, we show that our behavioural datasets can be similarly accounted for by a biologically-inspired population coding model, which describes how mappings between inputs and aspects of perception might be changed by adaptation. Importantly, this model does not encompass any metacognitive noise. In a third stage, we estimate what shape pre- and post-adaptation experiential distributions might have from modelling. These analyses suggest that experiential distributions could be both skewed, and have excess kurtosis, and that both of these outcomes can undermine SDT-based analyses of confidence.

The final phase of our study is the shortest, but in our minds most consequential. We show that the degree to which SDT-based analyses of confidence underestimate meta d' (relative to d') can scale with the degree to which experiential distributions are either skewed, or have excess kurtosis. The results of these analyses are not contingent on the veracity of the first three stages of our study, but are rather a direct computational consequence of SDT-based analyses of confidence being informed by probability distributions that are non-normally shaped. At the end of this section we explain why some popular SDT-based analyses of confidence are particularly susceptible to violations of the normality assumption, whereas other analyses informed by the SDT framework are robust against such violations.

Methods

Nineteen volunteers participated, 14 female, with a mean age of 23 (S.D. 2.6). All were experienced psychophysical observers by the end of the experiment, but almost all were novice at the beginning. Eighteen of the volunteers participated as partial fulfillment of the requirements of an undergraduate course, and the remaining participant is the third author of the study. 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. Participation involved ~11 hours of testing for each observer, split across 11 experimental sessions (usually conducted on different days).

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 100 individual black dots, subtending 0.02 degrees of visual angle (dva) in diameter at the retinae, presented within grey (luminance = 51 cd/m2) circular apertures, with a diameter of 5.5 dva (see Figure 2). The individual lifetime of each dot was 100ms, after which it was re-drawn at a random position within the aperture. Each dot was assigned a random initial age (between 0 and 100ms), so when there was no coherent movement in an RDK, dot updates created directionless flicker.

In the center of test displays there was a white (luminance = 102 cd/m2) static disc subtending 0.02 dva in diameter – which served as the fixation point. Participants were asked to fixate on this throughout each experimental session. The display was otherwise black. The 4 apertures that could contain RDKs were centered 6.9 dva to the left, right and above and below the central fixation point (see Figure 2).

Procedure

All trials had a matched sequence. They began with a 5 second adaptation period, in which two of four RDKs contained dots moving coherently in directions oppositely rotated from upward. The other two RDKs contained directionless flicker. There was then a 0.25 to 1.25 second inter-stimulus-interval (ISI), wherein no RDKs were presented, but the grey circular apertures could be seen. Two test RDKs, with dots also moving in oppositely rotated directions from upward, were then shown for 300ms. These were either presented in the same positions as the coherently moving RDKs (Adaptation trials), or in the same positions as flickering RDKs (Baseline trials). Participants were prompted to make a combined response, indicating which test had seemed to contain movement in a more clockwise direction (a categorical perceptual decision), and the degree of confidence they felt in this decision (by making a setting along a linear scale, with a minimal requirement of 5%, up to a maximum of 100%).

Participants were adapted to different magnitudes of direction difference, between the two adapting RDKs, in different experimental sessions. Direction differences of 0 to 360° were sampled, in steps of 40°, with left-side adapting stimulus movements rotated counter-clockwise from upward in steps of 20°, and right-side vice versa, such that at an adapting

direction difference of 360°, both adaptors contained downward motion. Each participant completed 10 adaptation conditions, in an individually randomized order.

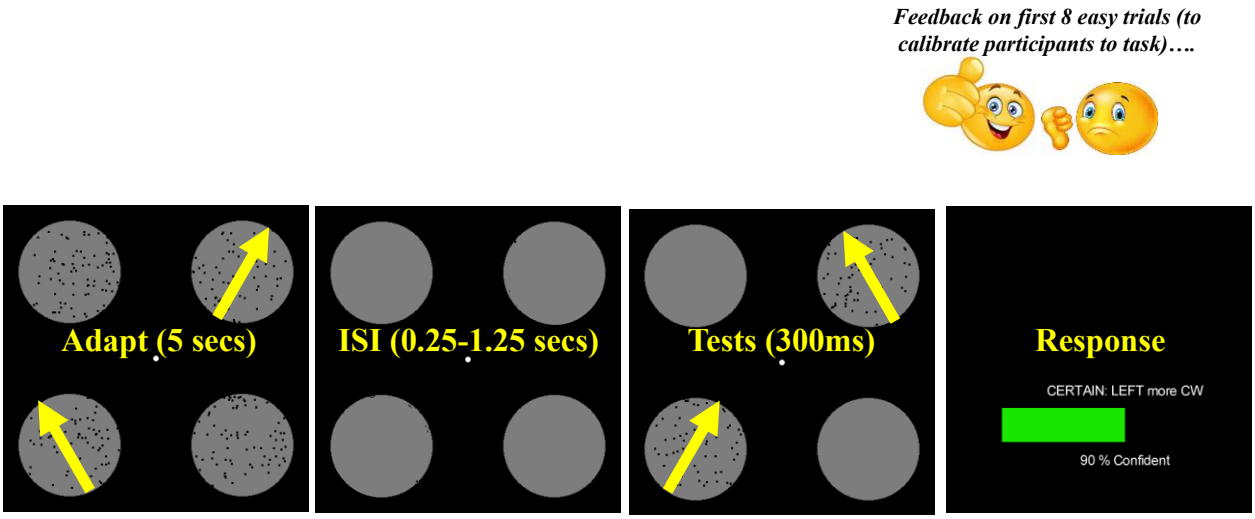

Figure 2. Graphic depicting the experimental protocol. On each trial participants were adapted (for 5 secs) to a pair of RDKs, diagonally positioned relative to the central fixation point (e.g. toward the bottom left and top right). Adapting motion directions were oppositely rotated from vertical, by one of a range of angular magnitudes. A single adapted direction difference prevailed throughout each block of trials. After variable inter-stimulus-intervals, participants viewed a brief pair of test directions, also oppositely rotated from vertical by one of a range of different angular magnitudes. Participants then simultaneously indicated which test had contained a more clockwise direction of motion, and the level of confidence they felt in this decision, by making a setting along a linear scale. Feedback was provided on the first 8 trivially easy trials, to (re)acquaint participants with the experimental task, while avoiding contaminating any intuitive insights into task performance, by training people to recognize when they had made correct or incorrect decisions about ambiguous inputs.

During each experimental session, the magnitudes of test direction differences were adjusted according to 1-up 1-down staircase procedures (Levitt, 1971), with decisions indicating that the right-side stimulus had moved in a more clockwise direction resulting in subsequent right-side tests moving in a more counter-clockwise direction, and vice versa. Four staircases were interleaved, two for each experimental condition (Baseline and Adaptation). One of these was initiated with right-side tests moving in a maximally clockwise rotated direction (given the range of test directions), and one was initiated with right-side tests moving in a maximally counter-clockwise direction. For most participants, maximal test directions were rotated +/-16° from upward, which was then adjusted in 2° steps. Some participants,

however, had more difficulty performing the task, so for these participants maximal test directions were rotated +/-32° from upward, and adjusted in 4° steps.

The first 8 trials of each experimental session were calibration trials, after which staircases were not updated. Feedback was only given, regarding the accuracy of perceptual decisions, on calibration trials. Test differences for calibration trials were set to the maximal possible test difference, to (re)familiarise participants with the task while avoiding contaminating intuitive insights into subsequent task performance, by training people to recognize when they had made a correct or incorrect decision about a perceptually ambiguous input. There were then an additional 100 trials for each experimental condition, all randomly interleaved – for a combined total of 208 individual trials in each experimental session. Staircases were re-set midway through experimental sessions, to ensure large test differences were sampled at least twice.

Individual data for each condition recorded during an experimental session were collated. Confidence ratings were then categorised, as high or low, relative to the median baseline confidence rating during that experimental session. Using a median split to categorize confidence settings (as low or high) was implemented to control for confidence bias. Our experience suggests that if you simply ask participants to rate their felt level of decisional confidence as low or high, a proportion of participants will tend only to report having one of these two levels of confidence – making it impossible to fit models to data. Allowing people to use a continuous scale to estimate felt levels of confidence, and then using a median split to categorize their confidence settings as high or low, controls for individual differences in

confidence bias. Doing this for each experimental session controls for confidence bias drift within a participant over time.

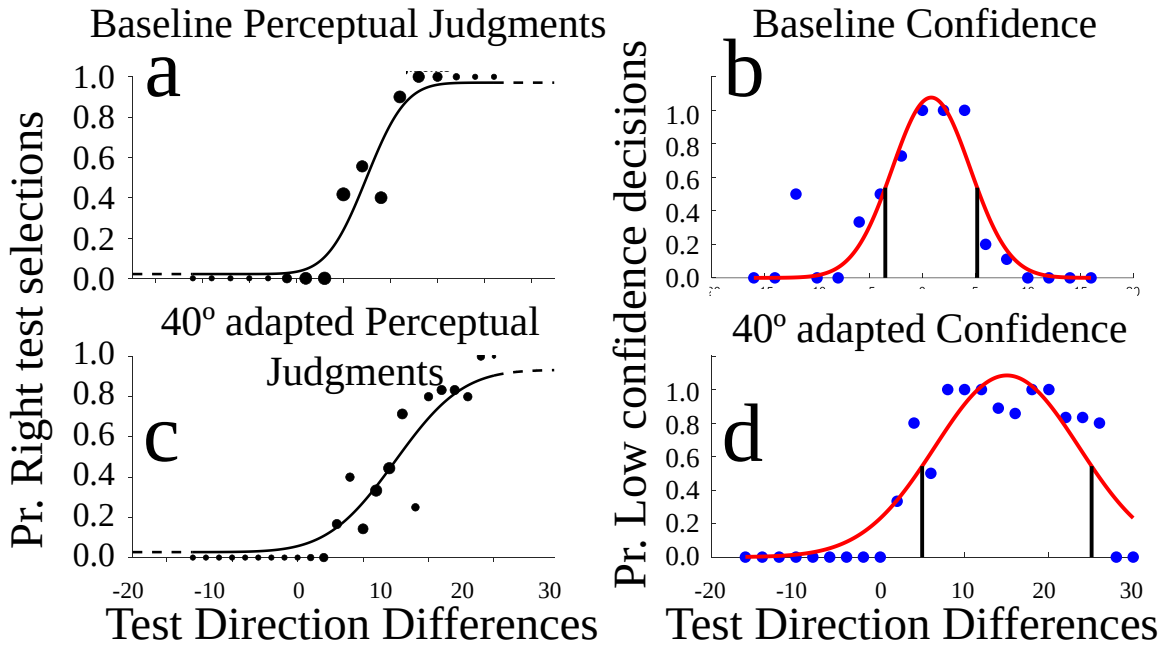

Figure 3. Sample participant datasets and function fits. This participant was selected for illustrative purposes, as their aftereffects for these conditions approximated the population average. **a)** Cumulative Gaussian function fit to data, describing the proportion of times that this participant selected the right-side test as having dots moving in a more clockwise rotated direction, as a function of test direction differences. These data are from unadapted baseline trials. Data point sizes scale indicate the number of trials completed for each test direction difference. **b)** Raised Gaussian function fit to data, describing the proportion of low-confidence decisions as a function of test direction differences. These data relate to the same unadapted baseline trials as depicted in Figure 3a. Data point sizes here are not scaled with numbers of trials. **c)** Details are as for Figure 3a, but these data relate to +40 degree adapted trials. **d)** Details are as for Figure 3b, but these data relate to +40 degree adapted trials.

A raised Gaussian function was fit to datasets describing low-confidence decisions as a function of test direction differences, for data from each experimental session (see Figure 3b & 3d). The peaks of these functions were taken as estimates of test directions Perceived to be Subjectively Equivalent (Confidence PSE estimates), and the Full-Width at the Half-Height (FWHH) of these functions were taken as estimates of the range of test directions that had elicited low-confidence decisions (a measure of the spread of uncertainty).

Cumulative Gaussian functions were fit to data describing the proportion of right-side tests seen as having moved in a more clockwise direction, as a function of physical test direction differences (see Figure 3a & 3c). The 50% point of these functions were taken as estimates of test directions Perceived as Subjectively Equal (Perception PSE estimates). Distances between 25 and 75% points of fitted functions were taken as estimates of the Just Noticeable Difference (JND) between test directions. Note that this last statistic is a standard measure of the precision of perceptual judgments.

Experimental sessions were usually completed on separate days, with breaks taken in-between sessions when multiple sessions were completed on the same day. Data informing individual aftereffect functions are generated from 2080 individual trials.

Data Availability / Pre-Registration

All data and analyses described in this paper are available as supplemental material, and will be made available via UQ eSpace. This study was not pre-registered.

Results

Stage 1: Behavioural Experimental Results

Adaptation-induced aftereffects were calculated from each experimental session, by subtracting adapted from unadapted summary statistics (i.e. adapted Perceptual PSEs – baseline Perceptual PSEs, etc…). Aftereffects expressed as a function of adapted direction differences, averaged across all participants, are depicted in Figure 5f-g. Adaptation to

directions rotated +/- ~20 to 100° from upward repelled perceived test directions away from adapted directions (see Figure 5f, blue data for Perceptual PSE aftereffects, red data for Confidence PSE aftereffects).

Adaptation to tests rotated +/-40° from upward produced an increase in Perceptual JNDs (plotted as negative values, i.e. baseline - adapted JNDs, as this is indicative of reduced precision). This is not obvious for other adaptors (see Figure 5g, blue data). Adaptation-induced reductions seemed greater in magnitude, and more widespread, for Confidence FWHHs (see Figure 5g, red data).

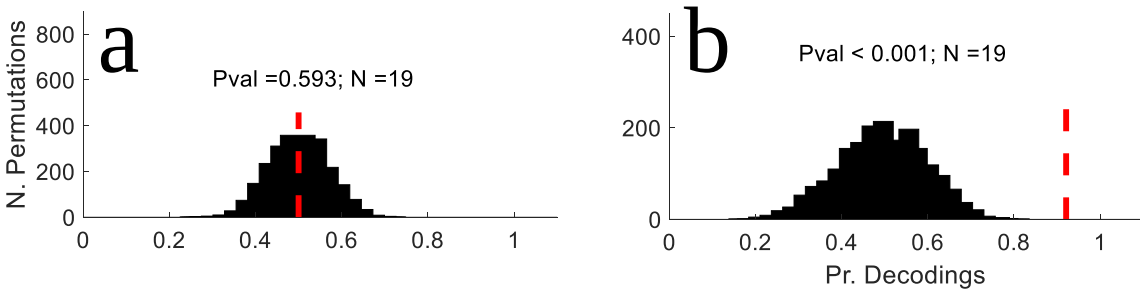

Figure 4. a) Histogram showing the numbers of simulations that have resulted in different proportions of nominally 'correct' classifications of aftereffect functions, each formed by randomly sampling from individual perception and confidence PSE aftereffect functions. This is a null distribution of chance classifications, that can be compared to our actual aftereffect function decoding process success rate (red dotted bar). **d)** Details are as for Figure 4a, but for a classification process for individual JND and FWHH aftereffect functions.

We conducted a non-parametric shuffle test, based on a nearest neighbour classification process with jack-knifed cross validation, as a formal test for a different impact of direction adaptation on Perceptual JND and Confidence FWHH aftereffects. Individual functions describing the impact of direction adaptation were classified as having described a JND or a FWHH aftereffect function based on similarities between that function and all other individual functions for these two categories. This process successfully classified 92% of individual aftereffect functions for these two categories (see Figure 4b). This success rate can be compared to classifications resulting from 2000 simulations, wherein data points from

each individuals' JND and FWHH aftereffect functions are randomly interchanged to form a function where there is no correspondence between the experimental condition and function data. These arbitrary functions are then classified according to the same process as the original procedure. The 2000 simulations provide a null distribution of chance classification success rates.

Comparing our actual classification success rate to this null distribution of chance classifications resulted in a p-value < 0.001 (see Figure 4b)—demonstrating that the success of our classification procedure was very unlikely to have emerged by chance. Overall, these data show that direction adaptation had a different impact on functions describing JND and FWHH aftereffects.

The last set of results can be compared to a matching set of analyses for our two sets of PSE changes—one calculated from categorical perceptual decisions, and one calculated from confidence judgements (from estimates of peak uncertainty). These analyses reveal that these two measures were interchangeable, with an actual decoding classification success rate of just 50%, and a non-parametric testing procedure resulting in a p-value of 0.593 (see Figure 4a). This demonstrates that categorical perceptual decisions and confidence judgements provide measures of perceptual central tendency that are equally impacted by direction adaptation. Scripts for these analyses are provided as Supplemental material code #4.

Stage 2: A biologically-inspired model to account for observed behavioural changes

To simultaneously account for the impact of adaptation on **1)** direction perception, **2)** the precision of direction judgments, and **3)** confidence in direction judgments, we created a labelled line observer model (Arnold et al., 2021; Clifford et al., 2001; Jin et al., 2005; Kohn

& Movshon, 2004; Kohn, 2007; Storrs & Arnold, 2012; Storrs & Arnold, 2015a,b). This assumes that sensory information is encoded as a pattern of responses to inputs from across a population of 'channels', each maximally responsive to a different direction (see Figure 5a-b). In our model the potential response of each channel to inputs is described by a normal distribution, with a standard deviation of 20°. Peak potential responses (channel tunings) are separated by 10°, ranging from ±180° (downward motion) to +170° in 10° steps – so our model has 36 channels.

The neural consequences of visual adaptation include reduced responding to inputs, and changes to both the optimal direction and to the range of directions that elicit responses (Kohn & Movshon, 2004). In our model these effects are operationalised by implementing a reduction in the response potential of model channels (up to a maximum of 95%) in proportion to how responsive each channel is to the adaptor in an unadapted state (Clifford et al., 2000; Jin et al., 2005; Kohn & Movshon, 2004; see Figure 5b), and by applying model channel tuning shifts (up to a maximum of 40°) away from adapted directions (Kohn & Movshon, 2004; see Figure 5b), also in proportion to how responsive each channel is the adaptor in an unadapted state. We operationalize trial-by-trial neural encoding noise by applying random changes (drawn from a uniform distribution ranging from 0 to 100%) to the 36 response potentials of all model channels (see Figure 3a-b; model code is provided as Supplemental material code #1).

Our model encodes a perceived direction value on each simulated trial by taking a sum of the direction labels of each channel, wherein each label is first multiplied by a product, of each channel's response (including noise) divided by the summed magnitude of all channel responses. Categorical perceptual decisions (i.e. is the stimulus moving clockwise or counter-

clockwise from vertical?) are decoded from adapted and unadapted (baseline) states of the model by comparing the weighted sum of all channel responses (the model response) to a criterion value (0°). Categorical confidence decisions (high / low) are decoded by indexing unsigned model response magnitudes against a criterion value (of 7°, so inputs decoded as moving > +/-7° from vertical elicit a high-confidence rating, and smaller decoded values elicit a low confidence rating). This confidence criterion was set to match half the average baseline Confidence FWHH (14°, S.D. 2.33), which equates to ~3x the average baseline Perceptual JND (4.5°, S.D. 1.29).

Model data from a simulated experiment are shown in Figure 5c-e. Repeated trials with identical physical inputs result in different encoded values, both at baseline (Figure 5c) and after adaptation to a +40° direction (Figure 5d), in each case due to random trial-by-trial changes in the response potentials of all model channels. Differences between values encoded by the baseline and by the +40° adapted model are also shown (Figure 5e). It is important to note that our model produces human-like performance when categorising inputs. In the simulated experiments cumulative Gaussian functions are fit to distributions describing the proportion of inputs encoded by the model as having moved in a clockwise direction relative to upwards (see Figure 4). The fact that these distributions describe probabilistic functions is due to the model, like humans, encoding a given input differently on repeated trials, due to trial-by-trial encoding noise.

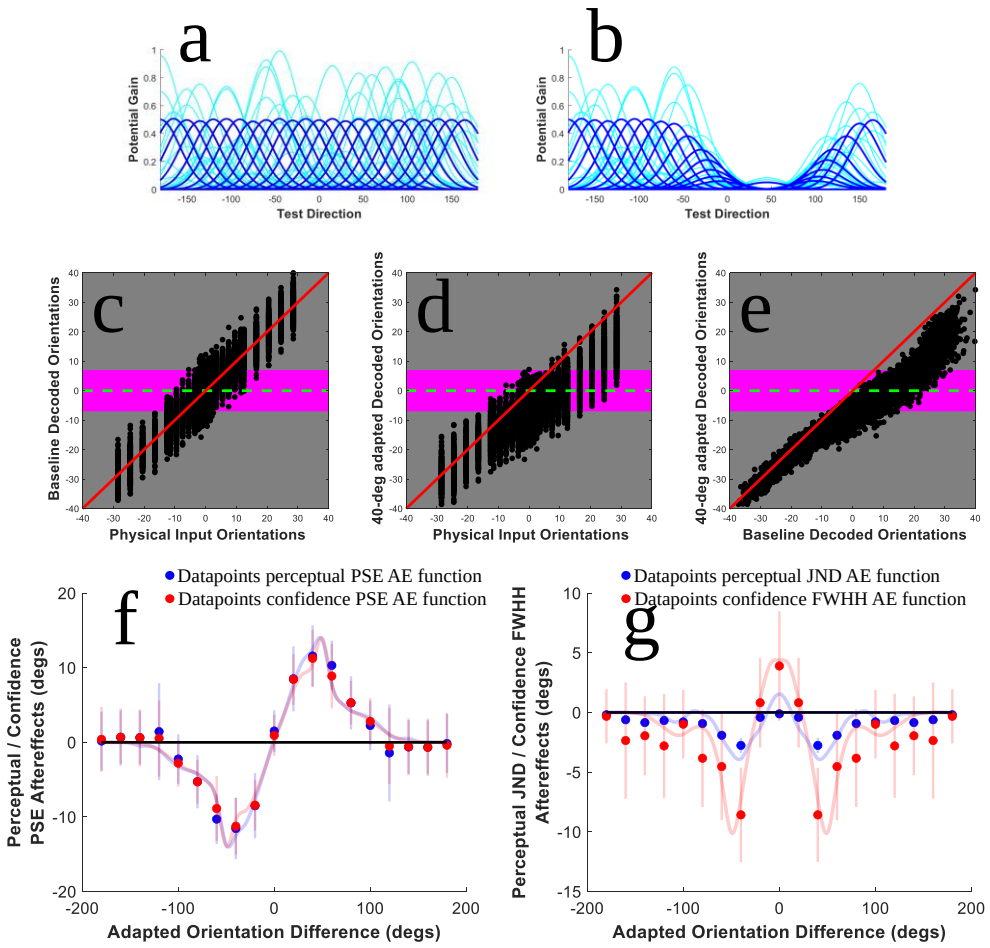

Figure 5. Depictions of **a)** unadapted (Baseline) model channels, and **b)** model channels adapted to a +40° direction. Faint blue lines depict response potentials on 3 simulated trials. Potential channel responses, averaged across 1584 trials, are also depicted (bold blue lines). **c)** X/Y scatter plot of encoded (Y axis) and physical (X axis) directions across 1584 simulated trials by the baseline model. Each physical input is encoded differently on discrete trials, due to simulated encoding noise, but on average the baseline model encodes inputs veridically, so datapoints cluster about the red oblique line (plotting veridical 1:1 input decodings). The horizontal green dotted line depicts the criterion value for perceptual categorizations (0°), and the edges of the horizontal pink rectangle depict the unsigned magnitude criterion for confidence categorizations (as low or high). **d)** Details are as for Figure 5c, but for the same simulated trials decoded by a +40° adapted model (with individual channels subject to the same nominal noise levels). **e)** Differences between perceived directions decoded by baseline (X axis) and +40° adapted (Y axis) models. **f)** Observer model fits (faint lines) to changes in both Perceptual (blue data) and Confidence PSEs (red data). Aftereffects following adaptation to negative direction differences are assumed to be mirror opposites of positive aftereffects. **g)** Observer model fits (faint lines) to changes in both Perceptual JNDs (blue data) and Confidence FWHHs (red data). Again, aftereffects following adaptation to *negative* direction differences are assumed to be mirror opposites of positive aftereffects (see Figure 5d). Model fits capture key qualitative features of aftereffects for all four derived dependent variables.

Our model provides a good account of the key qualitative features of all four behavioural aftereffect functions, describing the impact of adaptation on perceived direction (Figure 5f), on estimates of the precision of perceptual judgments (Figure 5g – blue data) and on the spread of uncertainty (Figure 5g – red data). All four functions have been fit simultaneously (i.e. with a common set of model parameters). These include: 36 model channels, response potentials normally distributed with standard deviations of 20°, channel tunings separated by 10°, ranging from ±180° to +170° in 10° steps, 95% maximum adaptation induced reduction of response potential (scaled to unadapted channel response to adaptor), 40° maximum channel tuning shift away from adapted direction (scaled to unadapted channel responses to adaptor), 7° confidence criterion.

We arrived at parameter settings for our model via a process of arbitrary but educated adjustments, to achieve a good simultaneous fit to all four of our aftereffect functions (which were in each case averaged across all participants). Given the high dimensionality of our model, we would not describe it in any way as optimal, or suggest it is superior to any related model. We would note, however, that other approaches to modelling often achieve a minimum of free parameters by assuming the existence of distributions of a specified shape (e.g. Locke et al., 2022; Maniscalco & Lau, 2016; Shekhar & Rahnev, 2021), and one thing our model demonstrates is that if the neural processes that might contribute to determining the shape of a distribution are made explicit, a host of assumptions might need to be implemented that are otherwise hidden. Regardless, we advance our model simply as an existence proof, that the key qualitative features of our four aftereffect functions can all be described by a biologically inspired population coding model. This motivated the next phase of our analyses, to see what shapes would describe the distributions encoded by our model for

repeated inputs, and to assess if any deviance from normality would undermine SDT-based analyses of confidence.

Stage 3: Does our model output non-normally shaped experiential distributions, to an extent that can undermine SDT-based analyses of confidence?

Stage 2 analyses modelled data for the baseline, and for each of the 10 adaptive conditions sampled in our behavioural experiment. For Stage 3 analyses, we will examine the products of a baseline and a +40° adapted model, as this adaptation condition was associated with the greatest adaptation-driven changes in our behavioural experiment. We simulated SDT experiments using our model to determine if repeated exposures to a given input, before and after adaptation, might result in non-normally-shaped distributions of different encodings, and to see if any deviance from normality would have a discernible impact on SDT-based analyses of confidence.

We simulated experiments for an unbiased observer, with 'signal' and 'noise' tests rotated by one of 101 different magnitudes (from +/- 1.5 to 2.7°) from a central perceptual category boundary – the direction encoded by the relevant model as upwards (~0° for our unadapted Baseline model, ~11.5° for our +40° adapted model). Confidence criteria were set to +/-7° from the central category boundary criterion (note that this confidence criterion is matched to the baseline performance of our human participants). This allows model encodings on each trial to be indexed against a central perceptual category boundary, and against confidence criteria to determine if a trial should be categorised as a noise presentation with high confidence, as a noise presentation with low confidence, as a signal presentation with low confidence, or as a signal presentation with high confidence (see Figure 1b).

This treatment of data allows us to implement a popular SDT-based analysis of confidence (Maniscalco & Lau, 2012) to arrive at d' and meta d' estimates (code for this analysis is provided as Supplemental material code #2). These analyses use the Matlab script fit_meta_d_MLE.m provided by Maniscalco & Lau (2012). Each simulated experiment and associated analyses relate to a single magnitude of signal / noise difference. Given that our model has no metacognitive bias (i.e. the same encodings inform both d' and meta d' calculations), if the model were producing normally-shaped distributions of encoded values, d' and meta d' estimates should fall along a line describing a 1:1 slope (marked by green oblique lines in Figure 6).

To compare model performance with categorisations based on normally-shaped distributions, we also included conditions where we randomly sample values from normally-shaped distributions. The mean values for these distributions were set to the 'signal' and 'noise' input values for the relevant simulated experiment (i.e. to one of a range of differences, from +/- 1.5 to 2.7° from the direction encoded by the relevant model as upwards). Standard deviations were set to the average JND of our human observers when making perceptual decisions for the same experimental condition (i.e. to 4.5° and to 6.7° respectively for the unadapted Baseline and for the +40° Adaptation conditions).

Results of stage 3 analyses are depicted in Figure 6. There are a few important features to note. First, SDT-based analyses informed by both of our model states (Baseline and +40° Adapted) underestimate meta d' relative to d' (see red data points in Figure 6a-b, and note that they all fall below the oblique green lines). This would normally be regarded as evidence for metacognitive insensitivity, with confidence judgments assumed to have been impacted by an additional source of noise relative to perceptual decisions (e.g. Fleming & Lau, 2014;

Maniscalco & Lau, 2012; Wixted & Stretch, 2004). In our simulated experiments we know this is not true. Second, note that both d' and meta d' estimates of sensitivity are reduced for our +40°adapted model, relative to our unadapted Baseline model, capturing the decline in sensitivity displayed by human participants post +40°adaptation (see Figure 5g).

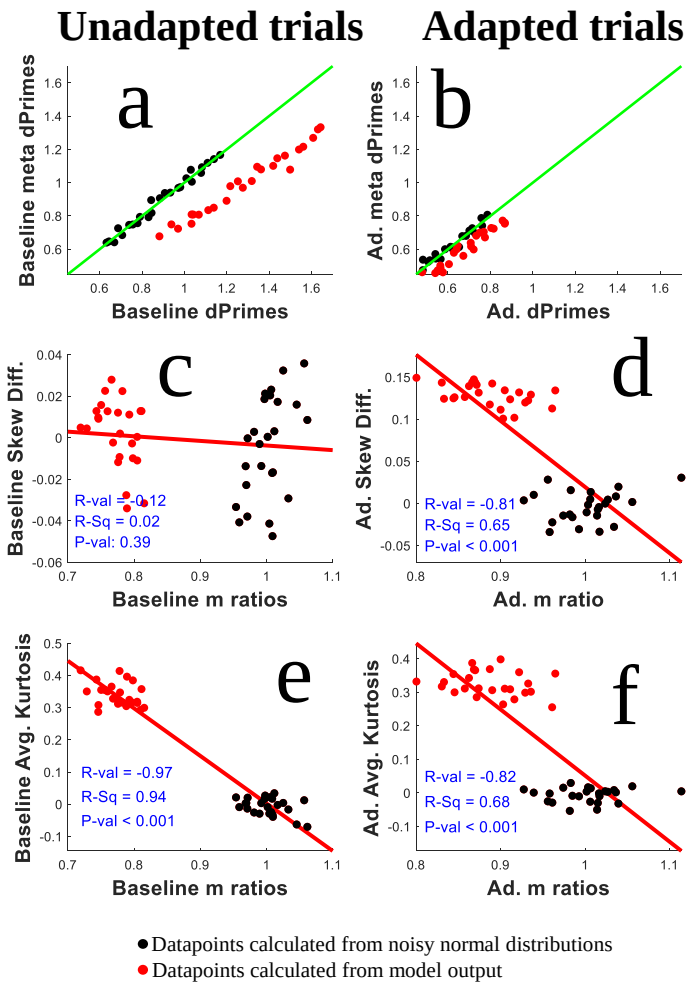

Figure 6. a) Scatterplot of d' (x-axis) and meta d' (y-axis) estimates calculated from our unadapted model output (red data points) or from values randomly sampled from normally-shaped distributions (black data). For each data point 50000 trials were simulated, 25000 non-target 'Noise' and 25000 target 'Signal' presentations. All d' and meta d' estimates should fall along the horizontal green line if they have been calculated from normally-shaped distributions. **b)** Details are as for Figure 6a, but for (red) data relating to a +40° adapted model. **c)** Scatter plot of Baseline meta d' : d' ratios (m ratios, X-axis) and the average skew of distributions of Baseline model encodings that have informed SDT-based analyses. **d)** Details are as for Figure 6c, but for data relating to +40° adapted models. **e)** Details are as for Figure 6c, but with average baseline distribution excess kurtosis on the Y-axis. **f)** Details are as for Figure 6e, but for data relating to +40° adapted models.

Readers should also note that while SDT-based analyses underestimate meta d' relative to d' when analyses are based on model encodings (see red data points, Figure 6a-b), estimates based on values randomly sampled from normal distributions (black data points) cluster about the green line marking a 1:1 ratio for meta d' : d' estimates – as predicted by SDT-based analyses of confidence. This can be regarded as a sanity check – signifying that SDT-based analyses of confidence perform as expected when analysis assumptions are met.

To diagnose the cause(s) of the underestimation of meta d' relative to d' in analyses informed by our model encodings, we have included scatterplots of meta d' : d' ratios (m ratios, X-axes) and differences in Signal and Noise skews (see Figure 6c-d, Y axes). We have also included scatterplots of m ratios (X axes) and excess Kurtosis (averaged across distributions describing both 'Signal' and 'Noise' test encodings, see Y-axes, Figure 6e-f). For unadapted Baseline model encodings, underestimation of meta d' relative to d' is predicted by excess kurtosis – by Signal and Noise distributions having a greater number of extreme encoded values than would be predicted by Normally-shaped distributions (see Figure 6e). There was no discernible relationship with a difference in baseline Signal and Noise distribution skews, as these distributions were effectively un-skewed (i.e. distribution skews were < +/- 0.02). For +40° adapted model encodings, underestimation of meta d' relative to d' was again predicted by excess kurtosis (see Figure 6f). However, for +40° adapted model encodings there was an additional robust relationship between underestimations of meta d' relative to d' and a difference between Signal and Noise distribution skews (see Figure 6d).

Our results to this point can be regarded as an existence proof, that if SDT-based analyses of confidence are applied to datasets that have been informed by non-normally-shaped experiential distributions (by Signal and Noise distributions that are either differently skewed,

or have excess kurtosis, or both), results can encourage flawed interpretations. How could researchers avoid drawing a false conclusion, that confidence has probably been adversely impacted by some additional source of noise relative to perceptual decisions (e.g. Fleming & Lau, 2014; Maniscalco & Lau, 2012; Wixted & Stretch, 2004), when data have instead been informed by a common source of information characterised by an non-normal shape?

Researchers often do not check for violations of the normality assumption when conducting SDT-based analyses. When they do, the most common check is to plot z-scored hit and false alarm rates, which can be done for target and noise inputs that elicit different levels of performance (e.g. Pastore & Scheirer, 1974) or for inputs that elicit different levels of confidence (e.g. Wixted & Stretch, 2004). When plotted, these should fall along a line with a slope of 1 if they have been calculated from normally-shaped experiential distributions with equal variance (Green & Swets, 1966). In our next set of simulations, we ask if researchers could reliably distinguish between datasets that have been informed by normally- and by non-normal distributions based on the slope associated with X/Y plots of z-scored hit and false alarm rates.

For this set of analyses, we conducted 1000 simulations wherein we randomly select 5 magnitudes of 'Signal' and 'Noise' test difference (from the 101 levels sampled in our second set of simulated experiments). For each condition we randomly select half the model encodings / values randomly sampled from normal distributions (i.e. data from 30000 simulated trials). From these data, we calculate best-fit linear slopes for sets of 3, 4 or 5 zscored hit and false alarm rates plotted on an X/Y scatterplot. From these, we calculate threshold criteria, for rejecting slopes as having likely resulted from analyses informed by

non-normally-shaped distributions (i.e. +/- 3 S.D.s from the average slope associated with zscored hit and false alarm rates for values randomly sampled from normal distributions).

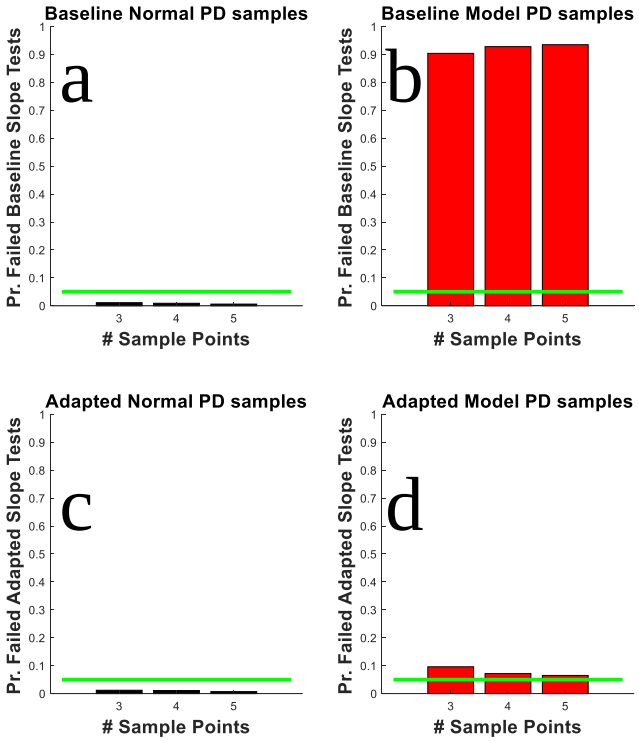

Figure 7. a) Bar plot depicting the proportion of 1000 simulations where fitted slopes were > +/-3 S.D.s from the average slope fit to z-scored Hit and False Alarm rates, calculated for random samples taken from normally-shaped distributions (i.e. failed slope tests). These data are circular (as average values and S.D.s were calculated from these data), so by definition there are only a small proportion of failed tests. **b)** Details are as for Figure 7a, but for slopes fit to z-scored Hit and False Alarm rates calculated from encodings by our unadapted Baseline model. **c)** Details are as for Figure 7a, but for data calculated for random samples taken from normally-shaped distributions with mean values and S.D.s set to mimic the performance of our human participants when adapted to +40°. **d)** Details are as for Figure 7b, but for data calculated from encodings by our +40° adapted model.

Bar plots depicting proportions of failed 'slope tests' are depicted in Figure 7. Note that by definition only a small number of slope tests are 'failed' when these criteria are applied to analyses informed by values randomly sampled from normal distributions (see Figure 7a and 7c). This is unsurprising, as these analyses are inherently circular (the test rejection criteria have been calculated from these same data). The point of these analyses is to calculate

rejection criterion values that can be applied to slopes calculated for z-scored model encodings. For our Baseline model (which underestimated meta d' by ~25%) this resulted in encouraging rejection rates (of ~95%). However, for our +40° adapted model (which underestimated meta d' by ~15%) this resulted in far smaller rejection rates (of ~5%). This means that experiential distributions that are sufficiently non-normal in shape to undermine SDT-based analyses of confidence *could not* be detected on ~95% of simulations on the basis of a z-scored slope tests, even when we had detailed knowledge of what tolerance we should adopt so as not to dismiss analyses that were actually informed by normally-shaped distributions (i.e. the average slope +/- 3 S.D.s, calculated from 1000 fits to pairs of 3, 4 or 5 z-scored Hit and False Alarm rates calculated from values randomly sampled from normal distributions).

Stage 4: What impact do different levels of distribution skew and excess kurtosis have on SDT-based analyses of confidence?

To this point our simulated analyses, describing the impact of non-normally-shaped distributions on SDT-based analyses of confidence, have related to analyses of model-generated encodings. This has a benefit, in that our model provides a good approximation of human performance in an actual experiment (see Figure 5), so the degrees of post-adaptation skew, and pre- and post-adaptation kurtosis, suggested by our model might resemble those produced by humans. However, this treatment limits analyses to 2 model states (Baseline and +40° adapted), that are likely characterised by a singular differential skew and average level of excess kurtosis.

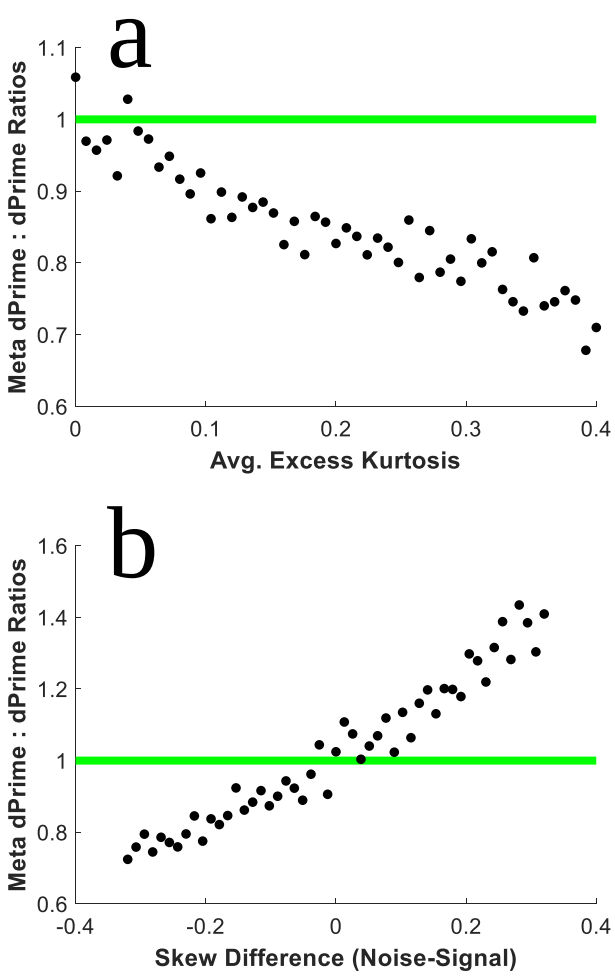

Figure 8. a) Scatter plot of meta d' : d' ratios (Y-axis) and the average excess kurtosis of signal and noise distributions (X-axis). As kurtosis levels increase from the level of kurtosis that is consistent with a normal distribution (i.e. 3), Meta d' estimates decrease as a ratio relative to d' estimates. **b)** Scatter plot of meta d' : d' ratios (Y-axis) and the differences in the skew of signal and noise distributions (X-axis). Meta d' estimates transition from exceeding d' estimates when Noise distributions are positively skewed and Signal distributions are negatively skewed, to being reduced relative to d' estimates when Noise distributions are negatively skewed and Signal distributions are positively skewed.

We decided to isolate and depict the effects of excess kurtosis and differential skew across a range of plausible values in a final set of simulations. For these, we selected the maximal average kurtosis produced by our model (~0.4) and the maximal level of a distributional skew (~0.16). To describe the effects of excess kurtosis, we used the Matlab pearsrnd command to create 51 pairs of distributions of signal and noise values with a common average offset (+/- 2° from a central perceptual category boundary of 0°), and a common standard deviation (of 3°), but with different levels of excess kurtosis – ranging from 0 to 0.4 (i.e. levels of kurtosis

range from 3, which describes a normal distribution, to 3.4). We use a confidence criterion value of 7 to classify generated values as noise presentations with high or low confidence, or as signal presentations with low or high confidence – and we estimate both d' and meta d' (code for this analysis is provided as Supplemental material code #3). As can be seen in Figure 8a, as excess kurtosis increases, so does the underestimation of meta d' relative to d'.

We conducted a final set of analyses, to describe the impact of differential signal and noise distribution skews. These are similar to our simulations describing the impact of excess kurtosis – but for these simulations we used the Matlab pearsrnd command to set excess kurtosis to 0 and Noise distributions to skews ranging from +0.16 to -0.16, and Signal distribution skews to a range from -0.16 to +0.16. As depicted in Figure 8b, when noise distributions are positively skewed and signal distributions are negatively skewed, meta d' tends to be overestimated relative to d', and when noise distributions are negatively skewed and signal distributions are positively skewed, meta d' tends to be underestimated relative to d'.

Discussion

Our behavioural experimental data described how direction perception, the precision of perceptual decisions, and confidence (i.e. the spread of uncertainty) are all impacted by adaptation to different direction differences (Stage 1 of our study; see Figure 5f-g datapoints and error bars). The key qualitative features of these changes could all be explained by a labelled-line population coding model, which describes how mappings between inputs, direction perception and confidence might all have been changed by adaptation (Stage 2 of our study, see Figure 5f-g smoothed lines). We reasoned that the success of this approach

could be challenging for SDT-based analyses of confidence, as these assume that repeated exposures to an input should elicit a normally-shaped distribution of different perceptual experiences (Green & Swets, 1966; Fleming & Lau, 2014; Maniscalco & Lau, 2012; Yarrow et al., 2011), whereas visual adaptation could encourage the emergence of non-normally-shaped distributions.

To assess the possibility that our model might encode non-normally-shaped distributions, in Stage 3 we conducted simulated experiments, to find out what shapes would describe the distributions of values encoded by our model, and more importantly to see if these would undermine SDT-based analyses of confidence. We found that distributions of encoded model values had excess kurtosis, both before and after adaptation (i.e. had more extreme values, relative to the predictions of a normal distribution), and were differently skewed post-adaptation (see Figure 6). Moreover, SDT-based analyses of confidence informed by these distributions were undermined – with meta d' tending to be systematically underestimated relative to d' – even though these two summary statistics were calculated from a common dataset with no metacognitive noise (see Figure 6).

In Stage 4 of our study, we conducted simulations to examine the impact of distribution skews and excess kurtosis on SDT-based analyses of confidence per se – independent of the operations of our model (see Figure 8). For these simulations we used simple matlab routines to generate data that conformed to differently shaped distributions, and quantified the effect these had on SDT-based analyses of confidence, which commit to the normality assumption (Fleming & Lau, 2014; Maniscalco & Lau, 2012). We established that these analyses are systematically undermined by either type of deviance from the normality assumption, with meta d' systematically over (when Signal distributions were positively skewed relatively to

Noise distributions; see Figure 8b) and underestimated (when Signal distributions were negatively skewed relative to Noise distributions, and when distributions had excess kurtosis; see Figure 8a-b) relative to estimates of d' – even though both estimates had been informed by a common dataset. Miyoshi et al (2022) have also recently demonstrated this adverse contingency, between excess kurtosis and the performance of SDT-based analyses of confidence. While our analyses were inspired by the first three stages of our study, this particular set of results does not rest on the veracity of the earlier stages of our study. The contingencies we have demonstrated in Stage 4 are factual, regardless of the motivations that led us to test for these.

Stage 4 analyses revealed that if SDT-based analyses of confidence are inadvertently informed by differentially skewed experiential distributions, meta d' can either be underestimated relative to d' (when Signal distributions are more Left-skewed) or overestimated (when Signal distributions are more Right-skewed, see Figure 8b). Results characterized by greater meta d' than d' estimates have been a source of some speculation as to probable cause (e.g. Rahnev & Fleming, 2019). Our results show that one possible cause is that SDT-based analyses of confidence have inadvertently been informed by skewed experiential distributions (see Figure 8b).

Our observer model is an unrealistic oversimplification – with some redeeming features

A reasonable criticism of our study is that our observer model is an oversimplification, that does not capture key features of decisional processes that are related to confidence. One glaring example is that we have made the simplifying assumption that all noise within the model resides within encoding stages. Other studies have enjoyed considerable success in explaining human behaviour, in studies of confidence, by describing how confidence criteria

might be subject to noise (e.g. Adler & Ma, 2018; Desender et al., 2022; Mamassian & de Gardelle, 2021; Miyoshi & Lau, 2020; Shekhar & Rahnev, 2021). Indeed, two of the authors have previously highlighted, in the context of time perception, that effects attributed to encoding changes can often be equally, or better explained by changes to decisional criteria (Yarrow et al., 2011). Our choice here, to assume a noiseless application of perceptual and confidence criteria, was not made because we think these are realistic or inconsequential assumptions, but rather because these are assumptions underlying popular SDT-based analyses of confidence (Fleming & Lau, 2014; Maniscalco & Lau, 2012) – and we ultimately wanted to subject our model data to this type of analysis. We note, however, that while it is necessary to assume fixed criteria when implementing SDT-based analyses, Maniscalco & Lau (2012) are well aware of the potential fallacy of this type of assumption.

What does this mean for SDT-based analyses of confidence?

Our data point to a dilemma – even in a context where we know that analyses of confidence have been informed by non-normally-shaped distributions, that fact could not reliably be detected by the standard procedure experimenters rely on to assess if a SDT-based analysis might have been informed by non-normally-shaped experiential distributions. This is to check for deviance from a 1:1 slope for z-scored hit and false alarm rates (Green & Swets, 1966; Pastore & Scheirer, 1974; Wixted & Stretch, 2004). This shows that when a SDT-based analysis of confidence suggests that a participant is metacognitively inefficient, there will be a degree of ambiguity – that rather than inefficiency, the meta d' - d' difference might instead be driven by analyses having been inadvertently informed by non-normally-shaped experiential distributions. So, is there any evidence that experiential distributions might be non-normally shaped in fact? We would argue that there is.

Evidence that experiential distributions might be non-normally shaped

It seems reasonable to presume that measures of perceptual sensitivity and precision will inversely scale with the magnitude of trial-by-trial changes in how inputs are encoded and experienced. Across many visual dimensions, sensitivity and precision are not uniform. Humans are more sensitive to direction and to orientation differences about cardinal (vertical and horizontal) as opposed to oblique angles (Appelle, 1972; Dakin et al., 2005; Girshick et al., 2011; Storrs & Arnold, 2015b). Spatial acuity scales with distance from fixation (Pollack & Mueller, 2005), and people are more sensitive to differences between slower speeds than to differences between faster speeds (e.g. Stocker & Simoncelli, 2006). In each of these cases, experiences triggered by repeated exposures to a given input might be associated with a greater range of experiences extending into regions of the perceptual dimension marked by less sensitivity/precision, and with a smaller range of experiences extending into regions associated with higher sensitivity/precision – in sum producing skewed experiential distributions. In human vision, skewed experiential distributions might be more the rule than an exception.

We believe there is also good evidence to suggest that human experiential distributions, from repeated exposures to a common input, might be characterized by excess kurtosis. Perceptual reproduction experiments (of perceived tilts and interval durations), suggest that human experiential distributions might be characterized by excess kurtosis (Acerbi et al., 2012; Anderson, 2014; Bays, 2016; Jabar & Anderson, 2015). In addition, the distribution of responses amongst cells tuned to different instances of a perceptual dimension in vision (i.e. to different orientations) is often marked by excess kurtosis – particularly for sections of a dimension 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). Non-normally shaped psychological dimensions may be more the rule than an exceptional circumstance.

SDT does not need to assume Gaussian-shaped experiential distributions – but assuming some other shape might not solve the problem

While popular instantiations of SDT commit to the normality assumption, including popular SDT-based analyses of confidence (e.g. Fleming & Lau, 2014; Maniscalco & Lau, 2012), SDT is overall agnostic about the precise shape of experiential distributions. Provided a particular shape of distribution can be assumed, areal overlaps can be estimated from behavioural data, and it does not matter if the distributions are assumed to have a normal or an non-normal shape (Green & Swets, 1966). Indeed, some researchers have advocated for performing SDT-based analyses of confidence using calculations that assume non-normally-shaped experiential distributions (e.g. Medha & Dobromir, 2021; Miyoshi et al., 2022; Winter & Peters, 2022), or which have a differently shaped distribution to either describe confidence (Boundy-Singer et al., 2022) or the variability of confidence criteria (Shekhar & Rahnev, 2021). The problem with this approach is that the criticisms we have made about the impact of erroneously assuming a normally-shaped experiential distribution would equally apply to erroneously assuming any other distribution shape, and we doubt the precise shape of distributions will ever be known in sufficient detail to preclude the possibility that results might differ from ground truth due to the wrong shape(s) having been assumed in calculations. Using this general approach, we believe there will always be some doubt as to whether evidence has revealed that confidence is truly subject to additional noise, or if the wrong shaped distributions have been assumed.

Non-normally-shaped distributions have a disproportionate impact on some SDT-based analyses of confidence. Other SDT-based analyses are robust. The problem is in the tails

A lot of our analyses have been informed by the SDT framework, and are based on calculations that assume normally-shaped distributions. For instance, both our analyses of behavioral data and our modelling have involved fitting cumulative Gaussian functions to datasets that describe categorical perceptual decisions (see Figure 3). As can be seen, this type of approach can result in good descriptions of data. So, why are many traditional analyses inspired by the SDT framework robust to violations of the normality assumption, whereas SDT-based analyses of confidence, which deliver meta d' / d' ratios (or differences), can be undermined?

The problem for some popular SDT-based analyses of confidence (Fleming & Lau, 2014; Maniscalco & Lau, 2012) is that type-2 confidence calculations are, in essence, an estimate of the areal overlap of the tails of target and non-target distributions (see Figure 1b), whereas standard type-1 SDT calculations are an estimate of the areal overlap of greater proportions of the same distributions (see Figure 1a). Moreover, the impact of any deviance from a normal shape, in terms of areal calculations (i.e. the numbers hits and false alarms, or the numbers of misses and correct rejections in signal detection parlance) is proportionally *exaggerated* for the tails of distributions, relative to any greater proportion of the same distributions. So, the proportional increase in numbers of hits and false alarms for type-2 high-confidence decisions from an increase in kurtosis (i.e. an increase in the fatness of distribution tails), will be *greater* than the proportional increase for type-1 SDT calculations (as these relate to greater proportions of the same distributions). This explains why standard SDT-based analyses are more robust to violations of the normality assumption than are SDT-based analyses of confidence.

We note that this dilemma will be exacerbated by a common practice in signal detection experiments, of adding a small value to each category of response when there are no trials for a given response category (e.g. when there are no high-confidence false alarms). In effect, this practice will impose an increase in the kurtosis of inferred experiential distributions (by artificially expanding the widths of inferred distributions, to encompass a category of response that was not encompassed in fact). As we have shown that excess kurtosis can lead to a false impression of metacognitive inefficiency, this practice seems ill advised if the experimenter is seeking an accurate estimate of metacognitive sensitivity.

The underlying problematic issue is that the impact of any deviance from a normal shape, in terms of areal calculations, will be disproportionate at the tails of experiential distributions relative to any greater proportion of the same distributions. Consequently, the problems we have identified will scale with the tightness of criteria used to de-mark low- form high-confidence decisions. If a very tight criterion is used (i.e. if participants require a large amount of encoded perceptual evidence to make a high-confidence decision, such as a relatively large perceived tilt), then high-confidence decisions will be restricted to a smaller range at the far extremities of experiential distributions, and the impact of any deviance from a normally shaped experiential distribution will be exaggerated. These issues would therefore be lessened if people instead adopt a more liberal criterion, and report a high-degree of confidence on more trials given smaller magnitudes of encoded perceptual evidence.

What's the solution?

For now, researchers should be mindful that popular SDT-based analyses of confidence (Fleming & Lau, 2014; Maniscalco & Lau, 2012) do not provide a ground truth measure of

human metacognitive sensitivity. They provide an approximation of what this might be, accompanied by some degree of unavoidable ambiguity, as to whether metacognitive sensitivity might really have differed, or if analyses might have inadvertently been informed by experiential distributions that are non-normally shaped.

To avoid interpretive ambiguity, estimates of metacognitive sensitivity should not be disproportionally impacted by any inadvertent difference between assumed and actual shapes of experiential distributions. A possible candidate is a recent modelling approach outlined by Mamassian & de Gardelle (2021). While this approach is informed by the SDT framework, it involves a modelling of the entirety of psychometric functions, and so it might be less susceptible to the influences we have identified. This, however, will need to be formally evaluated.

Until a better means of estimating metacognitive sensitivity can be developed and proven, researchers are urged to acknowledge ambiguities when interpreting the results of perceptual confidence experiments, to avoid over interpretations of data.

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

Adler, W.T. & Ma, W.J. (2018). Comparing Bayesian and non-Bayesian accounts of human confidence reports. PLoS Computational Biology, 14, e1006572.

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

Anderson, C. H. (1994). Basic elements of biological computational systems. International Journal of Modern Physics C, 5, 135–137.

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. 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. (2021).

Barrett, A.B., Dienes, Z. & Seth, A.L. (2013). Measures of Metacognition on Signal-Detection Theoretic Models. Psychological Methods, 18, 535 – 552.

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

Bouma, H. & Andriessen, J.J. (1968). Perceived orientation of isolated line segments. Vision Research, 8, 493 – 507.

Boundy-Singer Z.M., Ziemba C.M. & Goris R.L.T. (2022). Confidence as a noisy decision reliability estimate. bioRxiv https://doi.org/10.1101/2021.12.17.473249

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.

Clifford C.W.G., Wyatt A.M., Arnold D.H., Smith S.T. & Wenderoth P. (2001). Orthogonal adaptation improves orientation discrimination. Vision Research 41, 151-159.

Dakin, S.C., Mareschal, I. & Bex, P. J. (2005). An oblique effect for local motion: Psychophysics and natural movie statistics. Journal of Vision, 5(9) 878 – 887.

Desender K., Vermeylen L. & Verguts T. (2022). Dynamic influences on static measures of metacognition. Nature Communications 13, 4208.

Failor, S. W., Carandini, M., and Harris, K. D. (2021). Learning orthogonalizes visual cortical population codes. bioRxiv

Field D.J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. Journal of the Optical Society of America A 12, 2379 – 2394.

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.

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.

Jin D.Z., Dragoi V., Sur M. & Seung H.S. (2005). Tilt after-effect and adaptation-induced changes in orientation tuning in visual cortex. Journal of Neurophysiology U.S.A. 94, 4038-4050.

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.

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.

Kohn A. (2007). Visual Adaptation: Physiology, Mechanisms, and Functional Benefits. Journal of Neurophysiology, 97, 3155–3164.

Kohn A. & Movshon A.J. (2004). Adaptation changes the direction tuning of macaque MT neurons. Nature Neuroscience U.S.A. 7, 764-772.

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.

Locke S.M., Landy M.S. & Mamassian P. (2022). Suprathreshold perceptual decisions constrain models of confidence. PLoS Computational Biology 18(7): e1010318

Mamassian, P., & de Gardelle, V. (2021). Modeling perceptual confidence and the confidence forced-choice paradigm. Psychological Review, Advance online publication. https://doi.org/10.1037/rev0000312

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.

Medha, S. & Dobromir, R. (2021). The nature of metacognitive inefficiency in perceptual decision making. Psychological Review, 128, 45 – 70.

Miyoshi K. & Lau H. (2020). A decision-congruent heuristic gives superior metacognitive sensitivity under realistic variance assumptions. Psychological Review, 127, 655 - 671.

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.

Olshausen B. A. and Field D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381, 607–609.

Pastore, R.E. & Scheirer, C.J. (1974). Signal detection theory: Considerations for general application. Psychological Bulletin. 81(12), 945 – 958.

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.

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

Rouder J.N., Pratte M.S. & Morey, R.D. (2010). Latent mnemonic strengths are latent: A comment on Mickes, Wixted, and Wais (2007). Psychonomic Bulletin & Review, 17, 427 – 435.

Shekhar M. & Rahnev, D. (2021). The nature of metacognitive inefficiency in perceptual decision making. Psychological Review, 128, 45 – 70.

Stocker, A.A. & Simoncelli, E.P. (2006). Noise characteristics and prior expectations in human visual speed perception. Nature Neuroscience, 9, 578 – 585.

Storrs K.R. & Arnold D.H. (2012). Not all face after-effects are equal. Vision Research 64, 7-16.

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.

Tomassini, A., Morgan, M. J., & Solomon, J. A. (2010). Orientation uncertainty reduces perceived obliquity. Vision Research, 50, 541 - 547.

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

Winter, C.J. & Peters, M.A.K. (2022). Variance misperception under skewed empirical noise statistics explains overconfidence in the visual periphery. Attention, Perception & Psychophysics, 84, 1611 – 178.

Wixted, J.T. & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. Psychonomic Bulletin and Review, 11(4), 616-641.

Yarrow, K., Jahn, N., Durant, S. & Arnold, D. H. (2011). Shifts of criteria or neural timing? The assumptions underlying timing perception studies. Consciousness & Cognition20, 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.