American Psychological Association

Journal of Experimental Psychology: General

0096-3445/19/$12.00

2019, Vol. 148, No. 3, 437–452
http://dx.doi.org/10.1037/xge0000511

Sensory Noise Increases Metacognitive Efficiency

Ji Won Bang
Georgia Institute of Technology and New York University

Medha Shekhar and Dobromir Rahnev
Georgia Institute of Technology

Metacognitive efficiency quantifies people's ability to introspect into their own decision making relative to their ability to perform the primary task. Despite years of research, it is still unclear how visual metacognitive efficiency can be manipulated. Here, we show that a hierarchical model of confidence generation makes a counterintuitive prediction: Higher sensory noise should increase metacognitive efficiency. The reason for this is that hierarchical models assume that although the primary decision is corrupted only by sensory noise, the confidence judgment is corrupted by both sensory and metacognitive noise. Therefore, increasing sensory noise has a smaller negative influence on the confidence judgment than on the perceptual decision, resulting in increased metacognitive efficiency. To test this prediction, we used a perceptual learning paradigm to decrease sensory noise. In Experiment 1, 7 days of training led to a significant decrease in sensory noise and a corresponding decrease in metacognitive efficiency. Experiment 2 showed the same effect in a brief 97-trial learning for each of 2 different tasks. Finally, in Experiment 3, we combined increasingly dissimilar stimulus contrasts to create conditions with higher sensory noise and observed a corresponding increase in metacognitive efficiency. Our findings demonstrate the existence of a robust positive relationship between across-trial sensory noise and metacognitive efficiency. These results could not be captured by a standard model in which decision and confidence judgments are made based on the same underlying information. Thus, our study provides direct evidence for the existence of metacognitive noise that corrupts confidence but not the perceptual decision.

Keywords: perceptual decision making, confidence, metacognition, sensory noise, visual perceptual learning

Supplemental materials: http://dx.doi.org/10.1037/xge0000511.supp

When faced with difficult decisions, people not only make an informed choice but also can provide an estimate of the likelihood that their response was correct (Metcalfe & Shimamura, 1994). This judgment is usually provided in the form of a confidence rating (Mamassian, 2016). Confidence ratings are referred to as *metacognitive* judgments because they can be conceptualized as a second-order judgment about the accuracy of a first-order judgment (David, Bedford, Wiffen, & Gilleen, 2012; Fleming & Daw, 2017; Metcalfe & Shimamura, 1994; Yeung & Summerfield, 2012). The ability of confidence judgments to distinguish between correct and wrong answers determines the degree of visual metacognition. High metacognitive scores suggest that confidence judgments are informative and should be trusted, whereas low scores suggest the opposite. Despite the importance of understanding when confidence judgments are particularly useful and when they are less so, the factors determining the quality of metacognition are still not understood.

Metacognitive Efficiency

Research into the determinants of visual metacognition has been hampered by existing measures of metacognition. Traditional metrics include the area under the Type 2 curve (Fleming, Weil, Nagy, Dolan, & Rees, 2010), *phi* (the trial-to-trial Pearson correlation between confidence and accuracy; Nelson, 1984), and Type 2 d′ (Higham, Perfect, & Bruno, 2009). Such metrics are said to measure *metacognitive sensitivity* (Fleming & Lau, 2014): the ability of confidence ratings to predict accuracy (it should be noted that both *phi* and Type 2 d′ are not independent of criterion location and therefore should be avoided; Fleming & Lau, 2014). However, metacognitive sensitivity increases trivially as stimulus sensitivity increases (Maniscalco & Lau, 2012) and thus cannot be used to compare conditions for which stimulus sensitivity differs.

This article was published Online First November 1, 2018.

Ji Won Bang, School of Psychology, Georgia Institute of Technology, and Department of Ophthalmology, School of Medicine, New York University; Medha Shekhar and Dobromir Rahnev, School of Psychology, Georgia Institute of Technology.

This work was funded by a startup grant to Dobromir Rahnev from the Georgia Institute of Technology. These results have previously been presented at the following conferences: Cognitive Computational Neuroscience, New York, New York (2017), and Vision Science Society, St. Pete Beach, Florida (2018). All authors designed the studies. Testing and data collection were performed by Ji Won Bang and Medha Shekhar. All authors performed the data analysis and interpretation. Dobromir Rahnev drafted the manuscript, and Ji Won Bang and Medha Shekhar provided critical revisions. All authors approved the final version of the manuscript for submission. We thank Hakwan Lau, Brian Odegaard, and David Soto for helpful comments.

Correspondence concerning this article should be addressed to Dobromir Rahnev, School of Psychology, Georgia Institute of Technology, 654 Cherry Street Northwest, Atlanta, GA 30332. E-mail: drahnev@gmail.com

Recently, Maniscalco and Lau (2012) developed a way to measure *metacognitive efficiency* (Fleming & Lau, 2014): the quality of confidence ratings normalized by stimulus sensitivity. Their method computes an index (of metacognitive sensitivity) *meta-d′* that can then be divided by the level of stimulus sensitivity *d′*. The resulting metric is called M_{ratio} (Maniscalco & Lau, 2012, 2014). (Note that *meta-d′* can alternatively be normalized by subtracting *d′*; the resulting metric is called M_{diff}.) By constructing a measure of metacognitive efficiency, the development of M_{ratio} allows researchers to investigate metacognition independent of stimulus sensitivity.

Hierarchical Model of Confidence Generation

Armed with a measure of metacognitive efficiency, we explored what factors influence metacognitive efficiency and whether it is possible to manipulate it experimentally. To do so, we turned to existing models of confidence generation. Most current models assume that confidence is based on the exact same information used to make the perceptual decision (Fetsch, Kiani, Newsome, & Shadlen, 2014; Hangya, Sanders, & Kepecs, 2016; Pouget, Drugowitsch, & Kepecs, 2016; Rahnev, Bahdo, de Lange, & Lau, 2012; Sanders, Hangya, & Kepecs, 2016). These models predict that although higher stimulus sensitivity leads to higher metacognitive sensitivity, it results in constant metacognitive efficiency. However, several newer models have included an extra level of metacognitive noise that corrupts the confidence but not the decision judgments (De Martino, Fleming, Garrett, & Dolan, 2013; Jang, Wallsten, & Huber, 2012; Mueller & Weidemann, 2008; Rahnev, Nee, Riddle, Larson, & D'Esposito, 2016; Shekhar & Rahnev, 2018; van den Berg, Yoo, & Ma, 2017). We refer to these models as *hierarchical* models of confidence (Maniscalco & Lau, 2016) because they include two separate stages of noise corruption: The perceptual decision is corrupted by a first-level sensory noise, whereas the confidence rating is additionally corrupted by a second-level metacognitive noise (Figure 1A).

Because the perceptual decision and confidence are based on different information, hierarchical models of confidence allow in principle for dissociations between metacognitive and stimulus sensitivity resulting in nonconstant metacognitive efficiency (Rahnev & Denison, 2018). Indeed, many researchers have demon-

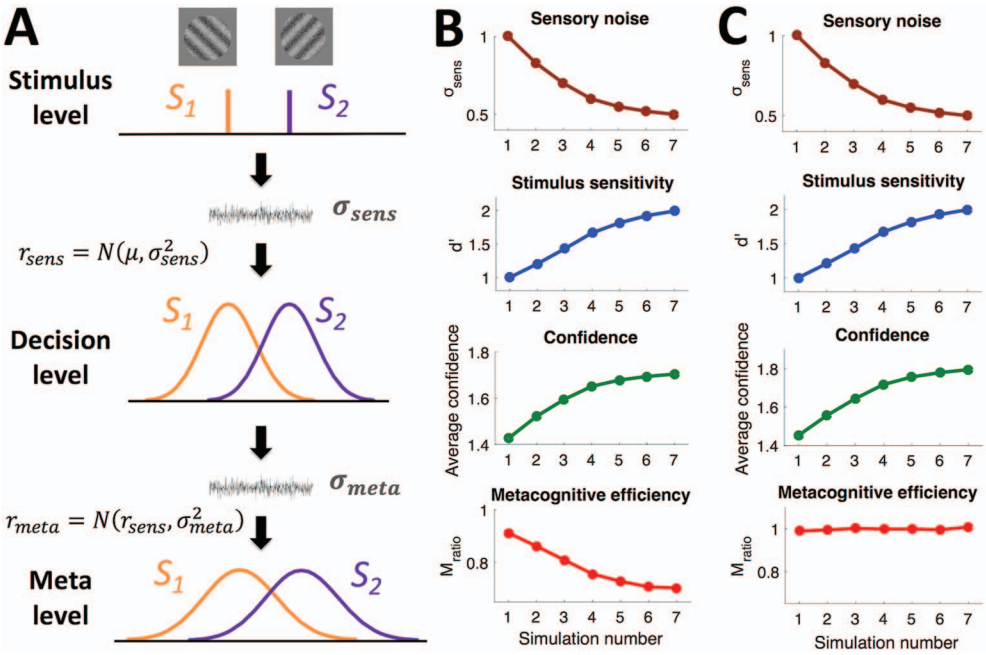

Figure 1. Hierarchical model of confidence. (A) Process model of confidence generation. At the stimulus level, two stimulus categories S_1 and S_2 (e.g., Gabor patches of counterclockwise and clockwise orientation) are presented. The stimuli are perfectly distinguishable. However, the internal representation at the decision level, r_{sens}, is corrupted by Gaussian noise, σ_{sens}, and thus the two stimulus categories are not perfectly distinguishable at the time of the decision. The confidence judgment is then made at the meta level based on an internal response, r_{meta}, that is derived from r_{sens} but is corrupted by additional noise, σ_{meta}. (B) Depiction of the model predictions. Seven simulations with a gradually decreasing level of sensory noise, σ_{sens}, show a gradual increase in sensory sensitivity *d′* and confidence ratings (given on a 2-point scale, such that high confidence is provided when probability of being correct exceeds 70%) but a decrease in metacognitive efficiency, M_{ratio} (for details on the simulations, see Method section of Experiment 3). (C) Depiction of predictions made by a standard model based on signal detection theory (SDT). The SDT-based model is equivalent to the hierarchical model but lacks a metacognitive noise stage. The same decrease in sensory noise leads to similar increases in stimulus sensitivity and confidence but no change in metacognitive efficiency. See the online article for the color version of this figure.

strated empirical dissociations between stimulus and metacognitive sensitivity (Fleming, Ryu, Golfinos, & Blackmon, 2014; Hauser et al., 2017; Maniscalco & Lau, 2015; Maniscalco, McCurdy, Odegaard, & Lau, 2017; Maniscalco, Peters, & Lau, 2016; Pleskac & Busemeyer, 2010; Rounis, Maniscalco, Rothwell, Passingham, & Lau, 2010). Nevertheless, this work was not generally aimed at testing predictions formally derived from hierarchical models of confidence.

Here, we report on a counterintuitive prediction of hierarchical models of confidence: Higher sensory noise should lower stimulus sensitivity but increase metacognitive efficiency. This prediction stems from the differential effect of sensory noise on stimulus and metacognitive sensitivity. Stimulus sensitivity is only corrupted by sensory noise, whereas metacognitive sensitivity is corrupted by *both* sensory and metacognitive noise. Therefore, increasing sensory noise is more detrimental to stimulus sensitivity than metacognitive sensitivity, resulting in higher metacognitive efficiency. Mathematically, within the context of hierarchical models of confidence, stimulus sensitivity d' can be expressed as the ratio of the signal and sensory noise, whereas metacognitive sensitivity *meta-d′* can be expressed as the ratio of the signal and a combination of sensory and metacognitive noise. Therefore, increasing sensory noise levels has a large negative effect on d' but a smaller negative effect on *meta-d′*, ultimately leading to an increase in their ratio (that is, M_{ratio}; see Figure 1B; for a detailed proof, see Method section of Experiment 1). This prediction holds regardless of whether the metacognitive noise is independent from or interacts with the sensory noise (see Supplementary Figure 1 of the online supplemental materials). Importantly, a standard model based on signal detection theory (SDT), which lacks a separate metacognitive noise stage, predicts that metacognitive efficiency remains constant for different sensory noise levels (Figure 1C).

Manipulating Sensory Noise

In order to test the prediction of the hierarchical model of confidence generation, one needs to find a way to manipulate the sensory noise. Importantly, the sensory noise in our model refers to the variability of internal evidence across trials. Note that there is substantial amount of work on the variability of the response of populations of neurons within a single trial (Haefner, Berkes, & Fiser, 2016; Ma, Beck, Latham, & Pouget, 2006; Orbán, Berkes, Fiser, & Lengyel, 2016), but our modeling approach does not work on the level of such within-trial neural variability. Instead, within SDT's framework, every trial is summarized with a single internal activity value (Macmillan & Creelman, 2005). It is the variability of these activity values across trials that we refer to as sensory noise.

Within the SDT framework, any manipulation that affects performance can affect the sensory noise, the sensory signal, or both. To change sensory noise is to increase or decrease the variability of the Gaussian distributions at the decision level, whereas to change the sensory signal is to increase or decrease the distance between the means of the distributions (Rahnev et al., 2011, 2013; Rahnev, Maniscalco, Luber, Lau, & Lisanby, 2012). A manipulation that decreases performance could do so by increasing the sensory noise or decreasing the sensory signal (or both), and it is impossible to know which one is affected just based on the performance change per se.

Therefore, in the current experiments, we employed two complementary strategies to affect sensory noise. First, we took advantage of the fact that training subjects on a perceptual task over many trials naturally leads to a decreased variability of the internal activity levels (Dosher & Lu, 1998, 1999, 2017; Petrov, Dosher, & Lu, 2005; Raiguel, Vogels, Mysore, & Orban, 2006). In other words, subjects exhibit smaller sensory noise later in the course of training as their sensory processing becomes more consistent and less variable. Second, we combined different ranges of contrast levels. The logic here is that combining more and more dissimilar contrasts in a single condition naturally increases the spread of internal activations and thus the sensory noise. Based on our hierarchical model of confidence, we can thus predict that these two manipulations will have opposite effects on metacognitive efficiency: Training should decrease sensory noise and therefore decrease metacognitive efficiency, whereas larger contrast ranges should increase sensory noise and therefore increase metacognitive efficiency.

Note that it is sometimes assumed that increasing stimulus noise (e.g., by adding random pixels noise to a Gabor patch) increases the sensory noise. Stimulus noise indeed makes subjects' estimates of the stimulus less precise, but the same is true if one is to decrease stimulus contrast (e.g., by reducing the contrast of the Gabor patch). It is generally unknown how such a decrease in precision in an estimation task translates into changes of the Gaussian distributions in a discrimination task. Within the SDT framework, higher stimulus noise and lower stimulus contrast could both decrease sensory signal, could both increase sensory noise, or could have different effects. Distinguishing between these possibilities is difficult and, to the best of our knowledge, has not been done before. In fact, we provide both intuitive (see Supplementary Figure 2 of the online supplemental materials) and simulation-based (see Supplementary Figure 3 of the online supplemental materials) arguments for why increased stimulus noise may not necessarily translate into increased sensory noise of the internal SDT distributions. Given the uncertainties regarding how manipulating stimulus characteristics affects SDT parameters, we do not include such manipulations in the present set of experiments.

Current Experiments

The current experiments tested the prediction that higher sensory noise increases metacognitive efficiency. In Experiment 1, we used a standard perceptual learning paradigm to decrease the level of sensory noise and observed a corresponding decrease in metacognitive efficiency. In Experiment 2, we applied the same logic but on a much finer time scale: We examined performance across a large group of subjects over a short training span. Just as in Experiment 1, we found that metacognitive efficiency, when computed across the group of subjects, decreased over the course of the training. Finally, in Experiment 3, we manipulated the level of sensory noise by using several ranges of contrast values and found that larger ranges increased metacognitive efficiency. Critically, a formal model comparison revealed that our hierarchical model provided a better fit to these data than a standard SDT-based model that lacks metacognitive noise. These results demonstrate that metacognitive efficiency depends on low-level stimulus character-

istics and provide strong support for the existence of metacognitive noise assumed by our hierarchical model of confidence.

Experiment 1

To test the counterintuitive prediction that decreasing sensory noise leads to lower metacognitive efficiency, we employed a perceptual learning paradigm. Twelve subjects participated in a 7-day training on a visual task. Subjects performed a 2IFC orientation detection task in which they indicated the interval (first or second) that contained a Gabor patch (Figure 2A). Stimulus intensity was adjusted using a two-down-one-up staircase procedure that allowed us to determine subjects' intensity threshold.

Method

Subjects. Twelve subjects participated in Experiment 1. The sample size was chosen in accordance with previous studies of perceptual learning and included a total of 84 days of testing (12 subjects coming for seven sessions each). All procedures were approved by the local institutional review board committee. Subjects reported normal or corrected-to-normal vision and provided informed consent.

Materials and Procedure. Subjects performed a two-interval forced-choice (2IFC) orientation detection task. Two stimuli were shown in quick succession and subjects indicated the interval (first or second) that contained the target. The target was a Gabor patch (circular diameter = 5°, standard deviation of Gaussian filter = 2.5°, spatial frequency = 1 cycle/degree, random spatial phase). We varied the stimulus intensity by substituting a random selection of the Gabor patch pixels with noise pixels, as done in previous experiments on perceptual learning (Seitz, Kim, & Watanabe, 2009; Shibata et al., 2017; Shibata, Watanabe, Sasaki, & Kawato, 2011). The noise pixels were generated using the formula $255 \times \frac{\sin(X)+1}{2}$, where X is a random variable with uniform distribution over the interval $[0, 2\pi]$. This method of generating random noise results in slightly higher variability in the noise values than choosing a uniform distribution over the interval [0, 255]. The stimulus intensity was defined as the percent of pixels that came from the original Gabor patch. For example, an intensity of 20% signifies that 20% of the pixels were selected from the Gabor patch

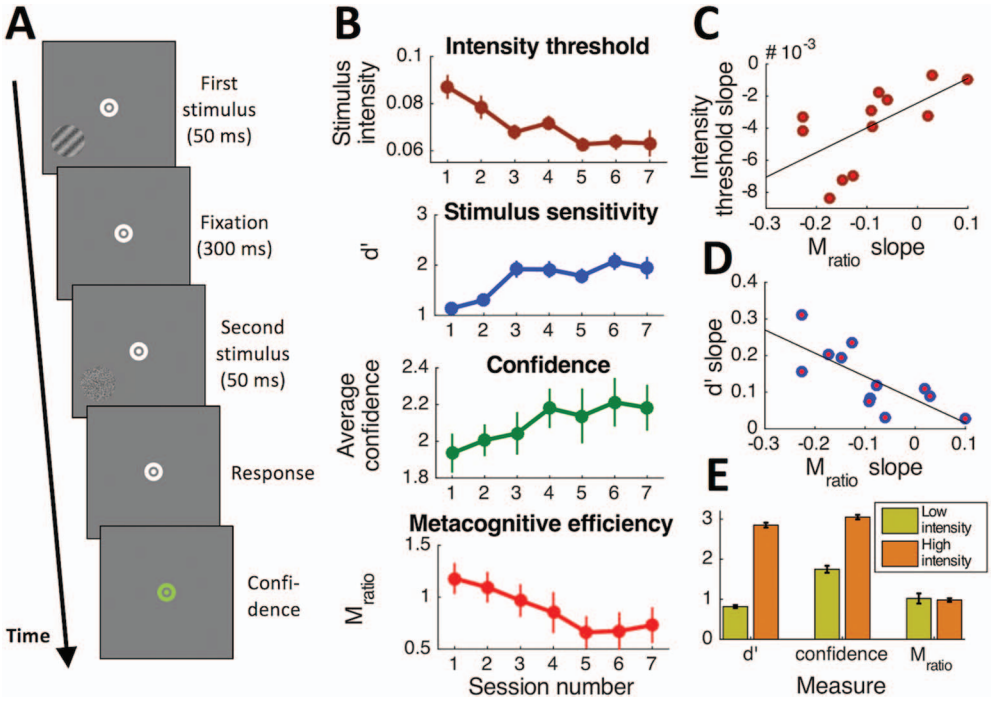

Figure 2. Visual training decreases metacognitive efficiency. (A) Subjects performed a two-interval forced-choice orientation detection task. Two stimuli—a target consisting of a noisy Gabor patch and a nontarget consisting of pure noise—were presented in a temporal sequence. Subjects indicated the interval in which the target appeared and provided a confidence rating on a 4-point scale. (B) Results of the 7 days of training indicate that intensity threshold gradually decreased. We further analyzed the stimulus intensity values in the 35th to 65th percentile range of each subject and found that stimulus sensitivity and confidence ratings increased. Critically, as predicted by our hierarchical model of confidence (Figure 1B), training decreased metacognitive efficiency. (C-D) The strength of the M_{ratio} decreases on a subject-by-subject basis depended on the decrease in intensity threshold (C) and increase in stimulus sensitivity (D). (E) Increased stimulus sensitivity does not automatically result in a decrease in M_{ratio}. Comparing low- and high-intensity stimuli (using a median split on all stimulus intensities) shows a large difference in stimulus sensitivity d' and confidence but no difference in M_{ratio}. Error bars indicate across-subject standard error of the mean. SDT = signal detection theory. See the online article for the color version of this figure.

image and the other 80% of the pixels were noise pixels. The nontarget consisted of noise pixels only (equivalent to 0% stimulus intensity). The target interval (first or second) was determined randomly on each trial.

Each trial started with a 500-ms fixation period. The two stimulus intervals lasted 50 ms each, separated by a 300-ms blank period. Subjects were asked to make two responses: first, to indicate the target interval, and second, to indicate their confidence level. Once the first response was made, the central fixation dot changed color from white to green to signal that the response had been recorded and to cue the need to make a second response. Subjects indicated their confidence using a 4-point scale.

We trained subjects on Gabor patches of a specific orientation (either 10° or 70°) presented in a specific visual quadrant (either lower left or lower right). The trained quadrant and orientation were determined randomly for each subject. Once an orientation and a quadrant were chosen for a specific subject, the training was done only on that orientation and in that quadrant. The center of the Gabor patch was positioned 4° away from the center of the screen in a direction of 45° from vertical so that it was located in the desired quadrant.

In addition to the trained condition, we included two untrained conditions. In the first untrained condition, the stimuli were presented in the trained quadrant but with the untrained orientation (either 70° or 10°, depending on which orientation was chosen for the training). In the second untrained condition, the stimulus was presented with the trained orientation but in the untrained quadrant (either lower right or lower left, depending on which quadrant was chosen for the training).

Subjects completed 12 blocks of trials per session. Each block consisted of trials in which a two-down-one-up staircase procedure continuously adjusted the stimulus intensity and terminated after 10 reversals. We the same procedure as in Shibata et al. (2017): The initial stimulus intensity was set at 30% in every block, with an adaptive step size that varied as a function of the current intensity (average step size = 2.4%, SD = 1%). Blocks consisted of an average of 42.1 trials and this number did not change over the course of the 7 days of testing, $F(6, 11) = 0.35$, $p = .91$. The intensity threshold for each block was calculated as the geometric mean of the last six reversals per block. In Sessions 2 to 6, all 12 blocks came from the trained condition, whereas in Sessions 1 and 7, four blocks were presented from each of the trained and two untrained conditions (in a randomized order). To keep the sessions as equivalent as possible, data analyses were performed on all four blocks from the trained condition in Sessions 1 and 7 as well as the first four blocks in Sessions 2 to 6. Analyzing all 12 blocks from Sessions 2 to 6 produced an equivalent pattern of results.

Stimuli were generated using Psychophysics Toolbox (Brainard, 1997) in MATLAB (MathWorks, Natick, MA) and were shown on a LCD display (1024 × 768 pixel resolution, 60-Hz refresh rate).

Analyses. To determine subjects' performance on the task, we computed the SDT measure d' (a measure of stimulus sensitivity) by calculating the hit rate (HR) and false alarm rate (FAR):

$$d' = \Phi^{-1}(HR) - \Phi^{-1}(FAR), \quad (1)$$

where Φ^{-1} is the inverse of the cumulative standard normal distribution that transforms HR and FAR into z scores. The measures of metacognitive efficiency M_{ratio} and M_{diff} were computed using the codes provided by Maniscalco and Lau (2012). Note that even though M_{ratio} is a ratio of two model-based (and thus relatively noisy) measures—*meta-d′* and d' these two measures are highly correlated and therefore their ratio is not necessarily noisier than the original measures. We determined the effects of training by computing the slope of change over the 7 experimental days using multiple regression. The exact functions that governed the changes over the seven sessions differed between subjects, and we fit a linear model because we were primarily interested in the overall trend of increase or decrease.

Prediction of hierarchical models of confidence. Our hierarchical model was built on the foundation provided by SDT (Green & Swets, 1966). Note that within the visual psychophysics tradition, confidence ratings are sometimes considered as a first-order judgment that operates directly on the sensory signal (Macmillan & Creelman, 2005). At the same time, outside of visual psychophysics, confidence ratings are often described to subjects and conceptualized by researchers as second-order judgment about the accuracy of a first-order judgment (David et al., 2012; Fleming & Daw, 2017; Metcalfe & Shimamura, 1994; Pouget et al., 2016; Yeung & Summerfield, 2012). Within this tradition, confidence judgments are thus typically referred to as *metacognitive* and we follow this terminology in the current paper. Here, we give a simple mathematical proof for why hierarchical models of confidence predict that higher sensory noise would lead to higher metacognitive efficiency. Stimulus sensitivity d' equals the ratio of the signal and noise present at the decision stage:

$$d' = \frac{\mu}{\sigma_{sens}}. \quad (2)$$

Equivalently, within the hierarchical model framework, metacognitive sensitivity *meta-d′* equals the ratio of the signal and noise present at the metacognitive stage. According to our hierarchical model of confidence, the signal at the metacognitive stage is still μ, but the noise is a combination of two Gaussian distributions with standard deviations of σ_{sens} and σ_{meta}. Therefore, we can derive that

$$meta\text{-}d' = \frac{\mu}{\sqrt{\sigma_{sens}^2 + \sigma_{meta}^2}}. \quad (3)$$

Combining Equations 2 and 3, we obtain

$$M_{ratio} = \frac{meta\text{-}d'}{d'} = \frac{\sigma_{sens}}{\sqrt{\sigma_{sens}^2 + \sigma_{meta}^2}}, \quad (4)$$

which, for a fixed σ_{meta}, is an increasing function of σ_{sens}. Therefore, as sensory noise σ_{sens} increases, so does metacognitive efficiency M_{ratio}. Note that Equation 4 does not feature the sensory signal μ, and therefore our model would predict that changing the sensory signal μ would have no effect on M_{ratio}.

Data and code availability. Data and codes for the analyses have been made freely available by the authors. They can be downloaded online at https://github.com/DobyRahnev/sensory_noise_metacognitive_efficiency.

Results and Discussion

Consistent with a decrease in sensory noise, training gradually decreased subjects' intensity threshold, $t(11) = -5.28$, $p = .0003$ (one-sample t test on the slope of change; Figure 2B). To compute

metacognitive efficiency M_{ratio}, we selected the same range of intensity values across all 7 days of training. We used intensity values in the 35th to 65th percentile range, but control analyses with larger percentile ranges (see the Supplementary Results of the online supplemental materials) or different stimulus intensity ranges chosen to equalize average intensity, the variability across the intensities, and the average number of trials per session (see Supplementary Figure 4) produced similar results. When considering only this range of intensity values, we observed that training increased stimulus sensitivity d', $t(11) = 5.2$, $p = .0003$ (Figure 2B) as well as average confidence, $t(11) = 2.43$, $p = .034$ (Figure 2B). Note that the confidence increase can be fully explained by the increase in stimulus sensitivity and does not by itself imply a change in metacognition (such change in confidence is predicted by both our hierarchical model and the SDT-based model in the absence of effects on metacognitive efficiency; see Figure 1B, C).

Critically, as predicted by our hierarchical model of confidence, the decreased sensory noise also resulted in decreased metacognitive efficiency M_{ratio}, $t(11) = -3.06$, $p = .011$ (Figure 2B). The same effect was also present for the alternative measure of metacognitive efficiency M_{diff} ($= meta\text{-}d' - d'$), $t(11) = -2.99$, $p = .012$. Note that although this effect was predicted by our hierarchical model (Figure 1B), it cannot be accounted for by a standard model with no metacognitive noise (Figure 1C).

Further, we examined whether the M_{ratio} decrease was indeed caused by the decrease in sensory noise or to some nonspecific effect of training. We found that subjects who showed a larger decrease in M_{ratio} also exhibited a larger decrease in intensity threshold ($r = .62$, $p = .03$; Figure 2C) and a larger increase in d' values ($r = -.74$, $p = .005$; Figure 2D), thus indicating that the M_{ratio} decrease is directly related to the change in performance on the perceptual task.

Further, one may worry that M_{ratio} has an intrinsic negative relationship with stimulus sensitivity d' or confidence level independent of sensory noise. To check for this possibility, we computed d', average confidence, and M_{ratio} across all seven sessions for the lower versus upper half of stimulus intensities used. As explained in the introduction (see also Supplementary Figures 2 and 3 of the online supplemental materials), by examining low- versus high-stimulus intensity, we are, in effect, comparing conditions that differ in sensory signal rather than sensory noise (in which case, our model predicts a constant M_{ratio}; see Method section of Experiment 1). We found that higher intensities led to a significantly higher d' (average $d' = 2.85$ and 0.82 for the upper and lower intensity halves, respectively), $t(11) = 46.23$, $p = 5.9 \times 10^{-14}$, and significantly higher confidence (average confidence = 3.05 and 1.75 for the upper and lower intensity halves, respectively), $t(11) = 26.12$, $p = 3 \times 10^{-11}$, but did not affect M_{ratio} (average $M_{ratio} = .98$ vs. 1.02 for the upper and lower intensity halves, respectively), $t(11) = -.38$, $p = .71$ (Figure 2E). Thus, the training-induced decrease in M_{ratio} cannot be explained as trivially arising from the corresponding d' or confidence increase: M_{ratio} appears to change only when the sensory noise is altered and not when the sensory signal is altered.

Finally, we examined whether training had any effects on the two untrained conditions, which were presented in Sessions 1 and 7 only. The untrained conditions involved presenting the trained orientation in an untrained quadrant (first untrained condition) or presenting an untrained orientation in the trained quadrant (second untrained condition). Comparing performance in Sessions 1 and 7 using paired t tests revealed no change in either d' (first untrained condition, $t[11] = .36$, $p = .73$; second untrained condition, $t[11] = -1.76$, $p = .11$) or M_{ratio} (first untrained condition, $t[11] = -.25$, $p = .81$; second untrained condition, $t[11] = .8$, $p = .44$). In other words, the effects of the training on both stimulus sensitivity and metacognitive efficiency were specific to the trained stimulus. These results provide further evidence against a direct effect of training on metacognition independent of its effects on sensory noise.

Experiment 2

Experiment 1 provided strong support for a causal link between decreased sensory noise and decreased metacognitive efficiency. It employed a standard perceptual learning design with extensive training over a number of days. In Experiment 2, we tested whether a much shorter learning period can also lead to decreased metacognitive efficiency. To this end, we recruited a large number of subjects ($N = 178$) to complete 97 trials of two different perceptual tasks. Critically, we inverted our analyses: Rather than combining many trials for each subject (the standard way of analyzing psychophysics data), we combined the data across subjects for a given trial (Figure 3A). This approach allowed us to track the evolution of across subject performance in terms of both stimulus sensitivity and metacognitive efficiency. Subjects engaged in coarse discrimination of low-contrast Gabor patch orientations (Figure 3B) and fine discrimination on high contrast Gabor patch orientations (Figure 3C).

Method

Subjects. Two hundred and one subjects participated in Experiment 2. The experiment was conducted online with subjects recruited using Amazon's Mechanical Turk. Subjects who performed at chance level or failed to clear our attention checks were excluded from the analyses. All procedures were approved by the local institutional review board committee. Subjects reported normal or corrected-to-normal vision and provided informed consent.

Materials and Procedure. Subjects performed two separate tasks—coarse and fine Gabor orientation discrimination. In each task, subjects discriminated between clockwise and counterclockwise oriented Gabor patches. In the coarse discrimination task, the stimulus was a Gabor patch of large tilt (±45°) overlaid on a noisy background composed of uniformly distributed intensity values. The overlaying was performed via pixel-by-pixel summation. In the fine discrimination task, the stimulus was a Gabor patch of small tilt (less than 1°) presented without any additional noise.

Each trial started with a fixation cross appearing at the center of the screen. The first trial of each block was preceded by a longer fixation period of 2 s to allow the subjects time to focus. All other trials had a variable fixation period that was sampled from a uniform distribution with a range of 300 to 700 ms. The stimulus was then presented for 500 ms. Once the Gabor patch disappeared, subjects were asked to make two responses using their keyboard: first, to indicate the orientation of the stimulus, and second, to rate their confidence on a 4-point scale.

We collected data from four batches of subjects. Three batches consisted of 50 subjects and one batch consisted of 51

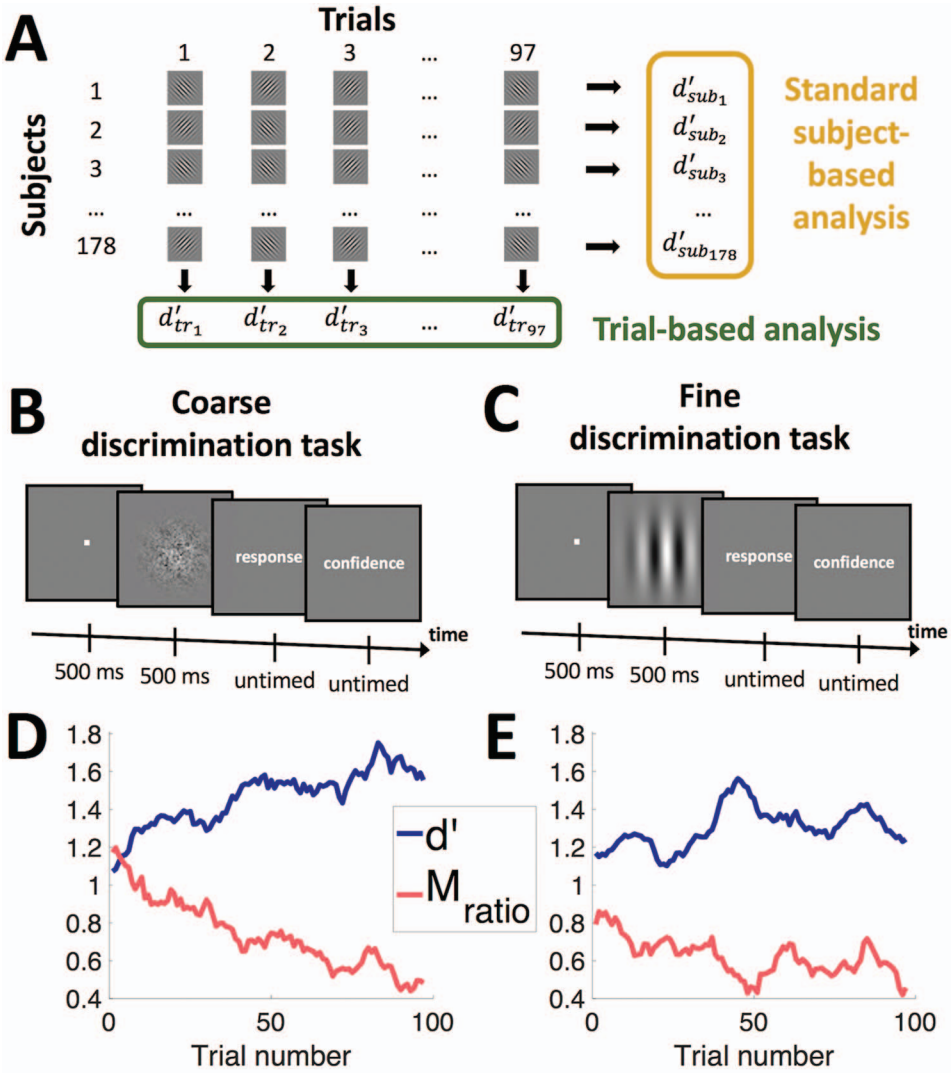

Figure 3. Visual training decreases across-subject metacognitive efficiency. (A) Depiction of standard subject-based analysis techniques (which consider all data for a given subject) and trial-based analysis techniques (which consider all data for a given trial number). We investigated the evolution of the trial-based d' and M_{ratio}. (B-C) Depictions of the two tasks. Subjects indicated the orientation (clockwise or counterclockwise from vertical) of a Gabor patch and provided a confidence rating on a 4-point scale. In the coarse discrimination task (B), the stimulus was a Gabor patch of low contrast but large tilt (±45°). In the fine discrimination task (C), the stimulus was a Gabor patch of high contrast but small tilt (<1°). (D-E) Practice resulted in a gradual increase in stimulus sensitivity d' but a decrease in M_{ratio}. Both of these effects were larger for the coarse (D) compared with the fine (E) discrimination task. The time courses are smoothed with a 11-point moving window for display purposes. See the online article for the color version of this figure.

subjects. In order to ensure similar average performance on both tasks, we varied the difficulty of each task across the batches (the difficulty for the first batch was determined by a separate pilot experiment with 20 subjects). For the coarse discrimination task, difficulty was manipulated by adjusting the contrast of the Gabor patch (subjects in the four batches experienced 5.5%, 6%, 5%, and 4.5% contrast, respectively). For the fine discrimination task, difficulty was manipulated by adjusting the offset from the vertical (subjects in the four batches experienced 0.62°, 0.7°, 0.7°, and 0.75° tilt, respectively). Average accuracies for each of the four batches were 76%, 87%, 76%, and 70% in the coarse discrimination task, and 70%, 77%, 77%, and 74% in the fine discrimination task. Overall, the percent of correct trials across all subjects was 76.44% for the coarse discrimination task and 74.12% for the fine discrimination task.

Subjects had to complete a total of 100 trials of each task. Each task was divided into five blocks of 20 trials each. Subjects were allowed to take breaks between each block, and the order of the tasks was randomized across subjects.

To ensure high data quality, we included six attention check trials—three in each task. These trials were designed to be much easier than the regular trials (contrast for coarse discrimination task = 15%, offset for the fine discrimination task = 5°), and subjects paying attention to the task were expected to have a high degree of accuracy for such trials. Therefore, we excluded subjects who responded incorrectly to more than two out of six catch trials (total = 15 subjects excluded). Additionally, we excluded subjects whose performance was close to chance level (<55% correct) on the noncatch trials of either task (additional eight subjects excluded). These criteria led to the exclusion of a total of 23 of the initial 201 subjects (11% exclusion rate). Note that the final analyses were based only on the 97 noncatch trials per task.

The Gabor stimuli were generated online via in-house code written in JavaScript and the experiment was designed using the jsPsych 5.0.3 library (de Leeuw, 2015). Subjects performed the experiment on their own computers outside the controlled environment of the laboratory. We attempted to minimize variability in the stimulus size in the following manner. To account for variability in the resolution and size of screens across computers, subjects were asked to adjust the size of images of real-life objects displayed on the computer screen to match their dimensions to the actual objects. Subjects were also asked to position the computer screen at an arm's distance (and were shown a picture of this configuration). This calibration was designed to ensure that the size of the stimulus displayed was uniform across different screens. Assuming an approximate arm length of 60 cm, the circular diameter of the stimulus was about 2° (potential variability caused by incomplete compliance with instructions likely made the actual visual angle was likely in the range between 1.7° and 2.3°). The jsPsych library also allowed us to obtain a precise reading of the actual duration of stimulus presentation, which was found to be 526 ms (SD = 17 ms).

Results and Discussion

As expected, stimulus sensitivity d' increased over the course of the 97 trials for both of our tasks (coarse discrimination task, $t[95] = 5.26$, $p = 8.8 \times 10^{-7}$; fine discrimination task, $t[95] = 2.34$, $p = .02$; t tests on the slope parameter in a linear regression; Figure 3D, E). Critically, as in Experiment 1, we observed a corresponding decrease in M_{ratio} (coarse discrimination task, $t[95] = -6.28$, $p = 9.9 \times 10^{-9}$; fine discrimination task, $t[95] = -2.31$, $p = .02$; t tests on the slope parameter in a linear regression; Figure 3D, E).

As can be seen in Figures 3D and 3E, the learning rate was different for the two tasks. Indeed, the d' increase was steeper for the coarse discrimination than for the fine discrimination task, $t(190) = 2.53$, $p = .01$. Importantly, we observed a corresponding effect in M_{ratio}, which showed a steeper decrease for the coarse than the fine discrimination task, $t(190) = -2.85$, $p = .005$, suggesting a direct relationship between the amount of learning and the decrease in metacognitive efficiency. All effects pertaining to M_{ratio} remained significant with the alternative measure of metacognitive efficiency M_{diff}. Finally, to further ensure that the effects on M_{ratio} were not an artifact of the way we performed the analyses, we simulated the effects of subject-level d' increase in the absence of any decrease in sensory noise and found no change in M_{ratio} (see Supplementary Figure 5 of the online supplemental materials).

Experiment 3

The results of Experiments 1 and 2 lend strong support for the notion that training-induced decrease in sensory noise leads to a corresponding decrease in metacognitive efficiency. Nevertheless, it remains possible that the results of both experiments depended on the use of training and that other manipulations of sensory noise would not produce equivalent results.

To investigate the influence of sensory noise independent of visual training, in Experiment 3, we manipulated the level of sensory noise directly. One straightforward strategy to increase the sensory noise is to construct increasingly larger ranges of contrasts (see Supplementary Figure 6 of the online supplemental materials). Indeed, combining more dissimilar contrasts together results in higher variability of difficulty levels and thus higher sensory noise. Twelve subjects performed a Gabor patch orientation discrimination task (Figure 4A) and completed 4,200 trials over the course of 3 testing days. The Gabor patches were presented with three different levels of contrast. To create different levels of sensory noise, we combined the three levels of contrast in different ways to construct four conditions that vary in the amount of sensory noise.

Specifically, in Level 1, we only considered a single contrast level at a time (lowest variability level). In Levels 2 to 4, we combined pairs of consecutive contrast levels (Contrast 1 Contrast 2; Contrast 2 Contrast 3), all three contrasts (Contrasts 1–3), or the most dissimilar contrasts (Contrast 1 Contrast 3), respectively (Figure 4B). By combining more and more dissimilar contrasts in the same analysis, we ensured that Levels 1 to 4 featured monotonically increasing sensory noise levels (see Supplementary Figure 6).

Method

Subjects. Twelve subjects participated in Experiment 3. We collected a total of 48 days of testing (12 subjects coming for four sessions each). The data from this experiment were already previously reported (Rahnev et al., 2013). All procedures were approved by the local institutional review board committee. Subjects reported normal or corrected-to-normal vision and provided informed consent.

Materials and Procedure. This study was originally reported as Experiment 2 in Rahnev et al. (2013). All study details can be found in the original publication. The subjects' task was to indicate the orientation (clockwise or counterclockwise) of a grating presented at fixation. Each trial began with 50-ms presentation of the grating followed by a fixation period of 200 ms. On each trial, the orientation of the grating was randomly selected to be tilted 10° clockwise or 10° counterclockwise away from vertical. The grating pattern was presented on an annulus (inner circle radius = 1.5°, outer circle radius = 4.5°). The stimulus consisted of a noisy background composed of uniformly distributed intensity values on top of which we overlaid a grating (0.5 cycles/degree) via pixel-by-pixel summation. Subjects were required to fixate on a small white square for the duration of the experiment. They were seated in a dim room 50 cm away from a computer monitor. Stimuli were generated using Psychophysics Toolbox (Brainard, 1997) in

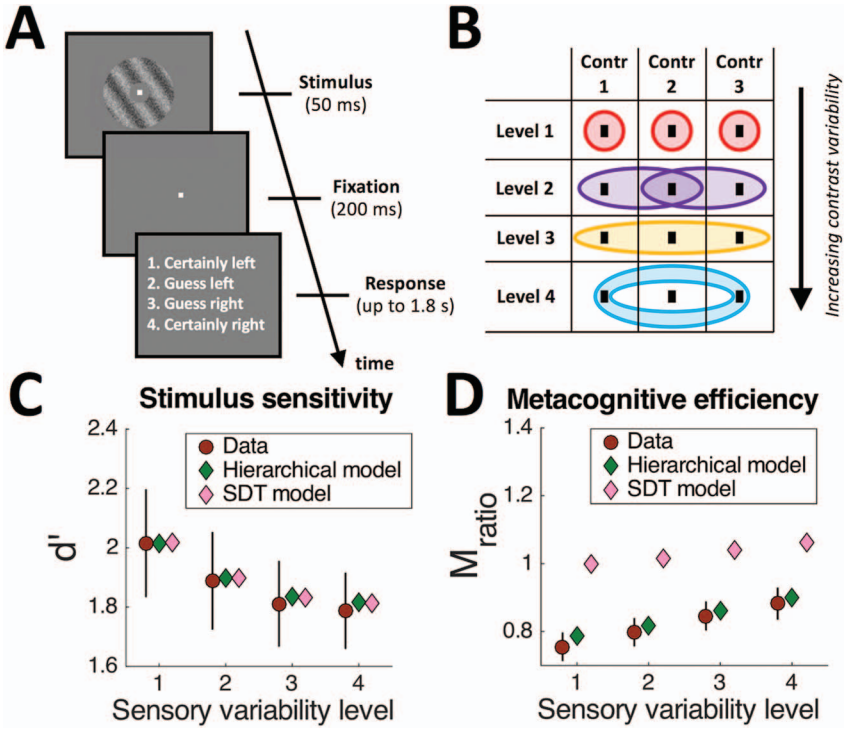

Figure 4. Experimentally increasing sensory noise increases metacognitive efficiency. (A) Subjects indicated the orientation (clockwise or counterclockwise from vertical) of a noisy Gabor patch and provided a confidence rating (on a 2-point scale) using a single button press. (B) Analysis logic: Three contrast levels were interleaved during the experiment. Different combinations of these contrasts resulted in different levels of stimulus variability. At the lowest level of variability (Level 1), each contrast was analyzed separately and the resulting d' and M_{ratio} values were averaged. At the next variability levels, increasingly disparate contrasts were combined: nearby contrast pairs in Level 2, all contrast levels in Level 3, and the far-contrasts pair in Level 4. The increased variability in stimulus contrast induced increased sensory variability (see Supplementary Figure 6 of the online supplemental materials). (C) The four levels of contrast variability were associated with a decreasing stimulus sensitivity d'. This effect was well captured by both our hierarchical model and a standard SDT-based model. (D) Higher stimulus variability led to higher metacognitive efficiency M_{ratio}. This effect was captured by our hierarchical model but not by the standard SDT-based model. Note that the SDT model predicts both higher M_{ratio} values and a shallower slope of M_{ratio} increase. Error bars indicate across-subject standard error of the mean. SDT = signal detection theory. See the online article for the color version of this figure.

MATLAB (MathWorks, Natick, MA) and were shown on a MacBook (13-in. monitor size, 1200 × 800 pixel resolution, 60-Hz refresh rate).

After each stimulus presentation, subjects used one of four keys to give their response indicating the perceived orientation of the grating and a wager on whether they were correct. Subjects used the keys 1 to 4, indicating "certainly left," "guess left," "guess right," and "certainly right," respectively. We note that although using two separate button presses for the perceptual decision and confidence is slightly more common, a very large number of experiments compress these into a single button press or, in the case of research with monkeys, a single saccade (Fetsch et al., 2014; Kiani & Shadlen, 2009; Lim, Wang, & Merfeld, 2017; Peters et al., 2017; Shekhar & Rahnev, 2018).

In order to facilitate the use of both low- and high-confidence ratings, we provided subjects with the following payoff structure. A correct "certain" (i.e., high confidence) choice was awarded with 2 points, whereas a correct "guess" (i.e., low confidence) choice was awarded with 1 point. An incorrect "guess" (i.e., low confidence) choice resulted in no points being won or lost, but an incorrect "certain" (i.e., high confidence) choice resulted in a loss of 2 points. We chose this point structure to ensure that subjects gave a sufficient number of both "guess" and "certain" responses. The optimal strategy for this payoff structure was to choose the "certain" choice only when the probability of being correct exceeded 66.7%. We informed subjects of this contingency in order to guarantee that all subjects were aware of the optimal strategy. Subjects had no trouble in understanding and using this reward structure. To further encourage optimal usage of the wagers, we gave the two subjects with the highest final scores an additional cash prize. Because the wagers that subjects used were a proxy for their confidence on each trial, we refer to the wagers as confidence ratings in the rest of the article.

Each trial lasted for 2 s. Subjects had 1.8 s to give their response after the onset of the stimulus. Once a response was given, the text indicating the four possible answers disappeared and the next trial

started. If a response was not given in the 1.8-s period, subjects were penalized by a subtraction of 4 points and the text was removed at the end of the 1.8-s period in order to avoid any potential interference with the processing of the stimulus in the next trial.

The study consisted of 4 days: 1 training and 3 testing days. In the initial training session on Day 1, subjects practiced with the task over the course of five blocks of 120 trials each. Based on the results of the training session on Day 1, we chose a grating contrast for each subject that would produce approximately 80% correct responses. However, we included two more levels of contrast: 75% and 125% of the chosen contrast. These three contrast levels were used on Days 2 to 4 without further adjustments, even if performance deviated from the 80% correct target for the middle contrast. Contrast level was chosen randomly on each trial and subjects were not explicitly informed about the presence of multiple contrast levels.

Days 2 to 4 involved three different conditions of theta burst stimulation (TBS): TBS to visual cortex, TBS to Pz, and sham TBS. Each of Days 2 to 4 started with five blocks of trials, followed by TBS administration, followed by another five blocks of trials. TBS to Pz and sham TBS had no effect on performance, whereas TBS to the visual cortex led to a slightly decreased sensitivity d' across the three contrast levels. Here, we combined all blocks from all three sessions regardless of TBS condition in order to increase the power of our analyses. Because we average over TBS conditions, our current analyses are orthogonal to the TBS effects. In each of Days 2 to 4, subjects completed a total of 10 blocks of 140 trials each for a total of 4,200 trials over the course of the 3 days. Note that the original publication excluded three of the subjects because they did not see phosphenes. These subjects were included here.

Model development. The model was fit only to the data in Experiment 3. The reason is that Experiment 3 featured a large number of trials per subject for each contrast level, which allowed us to precisely model the parameters of the internal response for each contrast level separately. On the other hand, in Experiment 1 subjects experienced a very large number of stimulus intensities, each presented only a few times, making it impossible to model precisely the internal response to each stimulus intensity. Experiment 2 featured an even bigger modeling challenge because the analyses were performed by considering the data from all subjects for a given trial: Fitting a model in this experiment would necessitate that the sensitivity of each individual subject on each individual trial is computed from a single trial at a time. Because of these considerations, we formally fit the model only to the data from Experiment 3.

Following standard assumptions dating back to the development of SDT (Green & Swets, 1966), each stimulus category was assumed to produce an internal response corrupted by Gaussian noise. Without loss of generality, we set one stimulus category to produce internal response $r_{sens} = N\left(-\frac{\mu}{2}, \sigma^2_{sens}\right)$ and the other stimulus category to produce internal response $r_{sens} = N\left(\frac{\mu}{2}, \sigma^2_{sens}\right)$, such that the distance between the two distributions was μ.

In Experiments 2 and 3, the stimulus categories were clockwise and counterclockwise orientations. However, in Experiment 1, we employed a 2IFC design, in which two stimuli—a target and a nontarget—were presented on every trial. To apply the model to that experiment, one can consider the first stimulus category to be the combination <target, nontarget> and the second stimulus category to be the combination <nontarget, target>. Then, by taking the difference between the internal responses to the first and second intervals, we obtain $r_{sens_{<\text{target, non-target}>}} = N(\mu_{target}, \sigma^2_{target}) - N(\mu_{non-target}, \sigma^2_{non-target}) = N(\mu_{target} - \mu_{non-target}, \sqrt{\sigma^2_{target}+\sigma^2_{non-target}})$ and, similarly, $r_{sens_{<\text{non-target,target}>}} = N(\mu_{non-target} - \mu_{target}, \sqrt{\sigma^2_{target}+\sigma^2_{non-target}})$. Then, the same equations apply after defining $\mu = 2(\mu_{target} - \mu_{non-target})$ and $\sigma^2_{sens} = \sqrt{\sigma^2_{target}+\sigma^2_{non-target}}$.

The perceptual decisions were modeled by specifying a decision criterion c_0 and confidence criteria $c_{-n}, c_{-n+1}, \ldots, c_{-1}, c_1, \ldots, c_{n-1}, c_n$, where n = number of confidence ratings. Importantly, the criteria $c_{-n}, c_{-n+1}, \ldots, c_n$ were constrained to be monotonically increasing with $c_{-n} = -\infty$ and $c_n = \infty$. Counterclockwise (clockwise) decisions were made based on whether the internal response r_{sens} was smaller (larger) than c_0. Confidence responses were given such that an internal response, r_{sens}, falling in the interval $[c_i, c_{i+1})$, resulted in a confidence of $i + 1$ when $i \geq 0$, and of $-i$ when $i \leq -1$ (where $i = -n, -(n-1), \ldots, n$).

The hierarchical model was constructed similarly but with the important addition of an extra layer of noise. The perceptual decision (about stimulus orientation) was made just as in the standard model. However, the confidence judgment was made on the internal signal at a metacognitive stage that was additionally corrupted by Gaussian noise with standard deviation of σ_{meta}, such that signal at the metacognitive stage was given by the formula $r_{meta} = N(r_{sens}, \sigma^2_{meta})$. Previous work using model comparison techniques supported the existence of such metacognitive noise and further suggested that this noise may remain constant across different stimulus conditions (Maniscalco & Lau, 2016). The confidence response was made equivalently to the standard SDT model. However, in rare cases, the model predicted that the subject would choose one stimulus category (e.g., "clockwise"), but the r_{meta} value would indicate that they should give a high confidence for the other stimulus category (e.g., "counterclockwise"). To avoid such inappropriate confidence ratings, confidence was constrained to always equal 1 when r_{meta} fell on the side of the decision criterion opposite to the category indicated by r_{sens}.

The seven simulations shown in Figures 1B and 1C were produced by setting σ_{sens} = 1, .83, .7, .6, .55, .52, or .5. All other parameters were kept constant: $\mu = 1$, $\sigma_{meta} = .3$ (in the hierarchical model) or 0 (in the standard SDT model). The criteria c_{-1}, c_0, and c_1 were set to values corresponding to 30%, 50%, and 70% posterior probability of a clockwise stimulus. Note that the pattern of results reported in Figure 1B and 1C is completely insensitive to the exact parameters chosen. We used simulations rather than direct numerical methods because computing directly M_{ratio} was made intractable by the fact that r_{sens} and r_{meta} sometimes fell on different sides of the decision criterion c_0.

As in previous studies employing a hierarchical model of confidence (De Martino et al., 2013; Jang et al., 2012; Maniscalco & Lau, 2016; Mueller & Weidemann, 2008; Rahnev et al., 2016; Shekhar & Rahnev, 2018; van den Berg et al., 2017), we chose to model the metacognitive noise as purely additive on top of the sensory noise. It is, however, possible that the metacognitive noise interacts with the sensory noise. The basic predictions of our model do not depend on the exact interaction of these two types of

noise and simulations of a model with metacognitive noise proportionate to the sensory noise (see Supplementary Figure 1 of the online supplemental materials) produced very similar results. Therefore, what is important here is the presence of metacognitive noise rather than its exact interactions with the sensory noise.

Model fitting. To model the effect of stimulus contrast in Experiment 3, we followed previous models (Qamar et al., 2013) and set $\sigma_{sens_{contrast(i)}} = C^{\alpha}$, where C was set to .75, 1, and 1.25 for the three levels of contrast (because contrast levels were 75%, 100%, and 125% of the subject-specific contrast threshold). Note that $\alpha > 0$ implies that σ_{sens} increases as a function of contrast, $\alpha < 0$ implies that σ_{sens} decreases as a function of contrast, and $\alpha = 0$ implies that σ_{sens} is equal for all contrasts. Thus, this modeling approach imposed a relatively minor constraint on the resulting σ_{sens} values for each contrast. Importantly, the parameter α was strongly correlated between the fits for the SDT and the hierarchical models ($r = .77$, $p = .004$), demonstrating that the superior fits of the hierarchical model were not related to an interaction between α and the extra parameter σ_{meta}.

Three of the 12 subjects exhibited M_{ratio} values larger than 1 in at least one condition. However, our model can only make the confidence ratings noisier than the perceptual decision and can therefore only predict M_{ratio} values smaller than or equal to 1. To allow a more precise fit of the M_{ratio} values, we included additional decision-level noise $\sigma_{decision}$ for these three subjects. This decision-level noise can be conceptualized as extra variability in the decision-level signal that does not propagate to the meta level. This variability makes it possible for d' to be smaller than *meta-d′* because it only affects the perceptual but not the confidence judgments. However, such decision-level noise cannot be fit independently from σ_{sens} and σ_{meta} because including this additional free parameter will make the model overparameterized: We will be using three free parameters to fit what are essentially two different quantities (d' and *meta-d′*). Therefore, for each of the three subjects, we simply set $\sigma_{decision}$ to equal the smallest value that allowed us to obtain M_{ratio} values that were as high as the ones observed for that subject (the three values were .2, .4, and .8). To avoid any bias, we applied this decision noise to both the hierarchical and SDT models. We note that our results remain the same if the three subjects who exhibited M_{ratio} values larger than 1 were simply excluded from the analyses.

The SDT and hierarchical models were instantiated with four and five free parameters, respectively. Importantly, the signal μ corresponding to each contrast level was not treated as a free parameter but was directly computed based on Equation 2 using the contrast-specific d' and sensory noise values. The standard SDT model thus had four free parameters: μ and the criteria c_{-1}, c_0, and c_1 (because confidence was provided on a 2-point scale). The hierarchical model was instantiated with five free parameters (the four from the SDT model and σ_{meta}). The criteria c_i were constrained to be nondecreasing and σ_{meta} was constrained to be ≥ 0.

We fit the models to the data as previously (Rahnev et al., 2011, 2013; Rahnev, Maniscalco, et al., 2012) using a maximum likelihood estimation approach. The models were fit to the full distribution of probabilities of each response type contingent on each stimulus type. Model fitting was done by finding the maximum-likelihood parameter values using a simulated annealing (Kirkpatrick, Gelatt, & Vecchi, 1983). Fitting was conducted separately for each subject's data by first running the fitting five times with a general starting parameter set, and then running the fitting five more times using a starting parameter set derived from the best fit from the previous stage. The best-fitting model from the second stage was used for further analyses. The Akaike information criterion (AIC) was used for model comparison, although the results remained the same if the Bayesian information criterion was used instead.

Results and Discussion

We found that higher levels of stimulus variability led to lower d', $t(11) = 4.53$, $p = .0009$ (Figure 4C). This result may appear surprising, because the different conditions consisted largely of the same actual trials that were simply combined in different ways. The robust but relatively modest decrease in d' can be explained by the nonlinear relationship between accuracy and d' (a detailed explanation can be found in Supplementary Figure 7 of the online supplemental materials). Indeed, both the hierarchical and the SDT-based models (see Figure 1B,C) could capture this decrease by simply modeling subjects' sensitivity to the three individual contrast levels (Figure 4C). Therefore, these d' results do not reflect changes in actual sensitivity between different ranges of contrast values and are not diagnostic in distinguishing between the hierarchical and the SDT models.

Critically, higher levels of sensory noise led to higher M_{ratio}, $t(11) = 6.21$, $p = .00007$ (Figure 4D; the same effect was observed for M_{diff}, $t[11] = 5.85$, $p = .0001$). This effect was quantitatively accounted for by our hierarchical model but not by the standard SDT model (Figure 4D). Most saliently, the SDT model predicted overall higher M_{ratio} values (average difference = 0.22), $t(11) = 6.06$, $p = .00008$. Note that even without metacognitive noise, the SDT model predicts increasing M_{ratio} values for higher levels of stimulus variability. The reason is that combining disparate contrast values results in violations of the Gaussian variability assumption, and this violation is greater for the higher variability levels. Nevertheless, the increase of M_{ratio} that can be attributed to violations of the Gaussian assumption is smaller than the increase in the data. Indeed, the SDT model predicted a shallower slope of increasing M_{ratio} values (.026 in model vs. .048 in data), $t(11) = 4.84$, $p = .0005$, indicating that metacognitive noise is needed to explain both the lower M_{ratio} values and the steep M_{ratio} increase caused by increased stimulus variability.

The results so far were obtained by analyzing the very same trials in different ways. This approach is not common and thus may appear contrived. Nevertheless, our design allowed us to perform a more traditional analysis in which every trial appears in only one condition. For that analysis, we compared the d' and M_{ratio} for Contrast 2 (low variability) with the d' and M_{ratio} for the combination of Contrasts 1 and 3 (high variability). Similar to the results from Figure 4D, we observed a significant increase in M_{ratio} for the high-variability condition, $t(11) = 3.59$, $p = .004$; see Supplementary Figure 8 of the online supplemental materials). Further, we confirmed that our results remain the same if we exclude three subjects who exhibited M_{ratio} values higher than 1 and for whom we had adjusted the models by adding a separate decision-level noise parameter (see Supplementary Figure 9).

Because the hierarchical model was more complex than the SDT model (it had one more free parameter), we compared the AIC for each model's fit. AIC measures the quality of the fit while punishing for the number of parameters. The hierarchical model still significantly outperformed the SDT model (average AIC difference across the 12 subjects = 23.48 signifying that the hierarchical model is 1.3×10^5 more likely than the SDT model). Model fits and AIC values for each subject are reported in Supplementary Table 1 of the online supplemental materials.

Importantly, as in Experiment 1, we confirmed that simply increasing d' does not necessarily lead to a decrease in M_{ratio} if the d' increase is caused by higher sensory signal rather than lower sensory noise. To demonstrate this point, we analyzed each level of contrast separately and found that higher contrast levels led to higher d' ($d'_{contrast1} = 1.06$, $d'_{contrast2} = 1.93$, $d'_{contrast1} = 3.21$; slope was significantly positive, $t[11] = 12.9$, $p = 5.5 \times 10^{-8}$) but did not significantly decrease M_{ratio} ($M_{ratio_{contrast1}} = .87$, $M_{ratio_{contrast2}} = .84$, $M_{ratio_{contrast3}} = .81$; slope was not different from zero, $t[11] = -1.04$, $p = .32$). Further, the d' increase from the lowest to highest contrast ($\Delta d' = 2.16$) was much higher than the increase from the lowest to highest variability level in Figure 4C ($\Delta d' = .25$), $t(11) = 16.49$, $p = 4.2 \times 10^{-9}$, indicating that the effects in Figure 4D cannot be simply the result of the difference in d'. Similar to Experiment 1, in which we compared low-versus high-stimulus intensities, the different contrast levels in Experiment 3 are likely to mostly influence the sensory signal rather than sensory noise. Therefore, this result is another confirmation of our model prediction that M_{ratio} varies as a function of the sensory noise but not as a function of the sensory signal.

Having confirmed that increasing the range of stimulus contrasts in Experiment 3 resulted in increased M_{ratio}, we looked for a similar effect in Experiment 1. We took advantage of the fact that Experiment 1 included a range of stimulus intensity values and examined the effect of selecting increasingly larger ranges of intensity values. We created four ranges (35th–65th, 25th–75th, 15th–85th, and 5th–95th percentile of all stimulus intensities used) and found that larger ranges did not change d', $t(11) = 1.53$, $p = .15$. but led to significantly higher M_{ratio} values (slope was significantly positive, $t[11] = 5.004$, $p = .0004$; see Supplementary Figure 10 of the online supplemental materials), thus mirroring the effects in Experiment 3. Therefore, manipulations of sensory noise based on learning or altered stimulus range resulted in equivalent effects on metacognitive efficiency across a variety of paradigms.

General Discussion

We found that sensory noise increases metacognitive efficiency. This effect was robust across experiments and manipulations. The increase of metacognitive efficiency with higher sensory noise was predicted by our hierarchical model of confidence generation that posits a stepwise organization of information flow for perceptual decisions and confidence. Conversely, a standard model based on SDT and lacking independent metacognitive noise could not explain our results. These findings demonstrate the possibility of directly manipulating subjects' metacognitive efficiency and provide strong evidence for the existence of metacognitive noise that corrupts confidence but not the perceptual decision.

A hierarchical model of confidence generation motivated our studies and provided excellent fit to the data. The model assumes that the information available for metacognition is corrupted by extra noise compared with the information available for the perceptual decision. Several previous articles have proposed similar architecture (De Martino et al., 2013; Jang et al., 2012; Maniscalco & Lau, 2016; Mueller & Weidemann, 2008; Rahnev et al., 2016; Shekhar & Rahnev, 2018; van den Berg et al., 2017). Here, we examined a strong, and previously unrecognized, prediction of hierarchical models on the relationship between sensory noise and metacognitive efficiency. Although previous work included metacognitive noise purely to improve model fit, we tested a direct prediction of hierarchical models. Therefore, our results provide some of the strongest evidence to date for the existence of independent metacognitive noise.

Although our results are consistent with the presence of second-level metacognitive noise, one may wonder whether they can be explained by alternative models. More specifically, several authors have advocated for dual-channel models, in which one (usually "conscious") channel influences both the stimulus decision and confidence, while another (usually "unconscious") channel only influences the stimulus decision (Del Cul, Dehaene, Reyes, Bravo, & Slachevsky, 2009; Jolij & Lamme, 2005). Such dissociation has been particularly prominent in explanations of the phenomenon of blindsight, in which above-chance performance seems to be accompanied with no subjective experience (Weiskrantz, 1996). Dual-channel models could potentially accommodate our learning results (Experiments 1 and 2) by postulating that training resulted in a small signal increase in the conscious channel (needed to explain the increase in confidence) and a larger signal increase in the unconscious channel (needed to decrease M_{ratio}). Critically, however, such models cannot explain the results of Experiment 3, in which larger contrast ranges resulted in higher metacognitive efficiency. In fact, it is unclear that any model devoid of metacognitive noise would predict both the learning and stimulus range effects reported here.

An important question concerns the source of this metacognitive noise. It is likely that confidence judgments are influenced by nonperceptual factors that do not contribute to the perceptual decision. For example, confidence ratings show strong serial dependence (Mueller & Weidemann, 2008) and can even be influenced by confidence ratings on a completely different task (Rahnev, Koizumi, McCurdy, D'Esposito, & Lau, 2015). Previous research has demonstrated that metacognitive efficiency is affected by fatigue (Maniscalco et al., 2017), working memory demands (Maniscalco & Lau, 2015), and heuristic use of perceptual evidence (Maniscalco et al., 2016), and can be enhanced pharmacologically via noradrenaline blockade (Hauser et al., 2017). Further work has shown that metacognitive—but not sensory sensitivity—can be affected by transcranial magnetic stimulation (Rahnev et al., 2016; Rounis et al., 2010; Ryals, Rogers, Gross, Polnaszek, & Voss, 2016; Shekhar & Rahnev, 2018) or lesions (Fleming et al., 2014) to the prefrontal cortex. Note that all of these factors affecting metacognitive noise are independent of whether confidence ratings are given simultaneously with the perceptual decision (as in Experiment 3) or after it (as in Experiments 1 and 2).

We modeled the metacognitive noise as affecting the signal on which the confidence judgments are made (see Figure 1A). However, an alternative possibility is to model the metacognitive noise as affecting the confidence criteria rather than the signal itself. In fact, in the absence of additional manipulations, these two con-

ceptualizations of metacognitive noise are equivalent and therefore cannot be distinguished from each other. We remain agnostic about the relative contributions of signal versus criterion variability to the metacognitive noise.

Given that we manipulated low-level stimulus characteristics, a critical question concerns whether subjects' metacognitive ability was truly altered. To clarify this issue, it is important to make a distinction between a subject's intrinsic capacities and the actual trustworthiness of her confidence judgments. These two concepts are typically related but become dissociated in certain situations. For example, as we pointed out in the introduction, assuming the same intrinsic capacity, increasing d' makes metacognitive judgments more predictive of one's accuracy. Similarly, we do not think that changing low-level stimulus characteristics (at least the manipulations in our studies) leads to a change in the quality of the downstream metacognitive processes. In other words, the intrinsic ability of the subjects is likely unaltered. However, higher sensory noise makes confidence ratings more predictive of one's accuracy relative to what one would expected for this level of accuracy (i.e., sensory noise increases metacognitive efficiency). In other words, although we do not think that higher sensory noise affects subjects' intrinsic metacognitive capacity, it does improve the quality of their confidence ratings as measured by metacognitive efficiency.

Our finding of a positive relationship between sensory noise and metacognitive efficiency raises the question as to how metacognitive scores should be interpreted. Influential theories pose that metacognition stems from second-order monitoring processes (Shimamura, 2000) that can be temporally separated from the first-order perceptual decision (Pleskac & Busemeyer, 2010). The contents of these second-order metacognitive processes are often assumed to reflect the contents of consciousness (Kunimoto, Miller, & Pashler, 2001; Persaud et al., 2011). However, our results demonstrate that although metacognitive judgments may indeed be related to consciousness, they cannot generally be used as a direct measure of consciousness (Jachs, Blanco, Grantham-Hill, & Soto, 2015). Indeed, perceptual learning has been argued to increase consciousness (Schwiedrzik, Singer, & Melloni, 2011) but, as seen here, decreases metacognitive efficiency. We see metacognitive scores as invaluable in constructing and testing models of decision making but remain agnostic about their relationship to constructs such as consciousness and working memory.

What is the most straightforward way for future studies to induce increased levels of sensory noise in order to test further our hierarchical model? It appears that the easiest way is to simply compare conditions of increasingly dissimilar contrast levels. Nevertheless, researchers would need to keep one caveat in mind: Combining more than one contrast level produces non-Gaussian distributions and therefore violates SDT's assumptions. In Experiment 3, we accounted for such deviations by modeling each individual contrast separately. Failing to do so may lead to inaccurate d' values. Future research should explore new methods for increasing sensory noise without violating SDT's Gaussian variability assumption.

A potential limitation of our Experiment 2 is that data were collected online without the usual tight control present in the laboratory. It should be noted that the effects of many classical perceptual and cognitive tasks have been robustly replicated in online studies (e.g., Semmelmann & Weigelt, 2017). Further, the most severe concerns about online experiments have to do with correlational and across-subject designs. In our case, the design was purely within subject. In other words, subjects acted as controls for themselves, and therefore issues like noncompliance or partial compliance, biased sampling, and so forth could not affect our results.

We tested the hierarchical model of confidence in the context of perceptual decision making, but our framework is general and we expect that the same model will apply to other modalities as well. For example, hierarchical models have already been proposed for confidence ratings related to items within working memory (van den Berg et al., 2017). Future research should test the predictions of our model in other domains such as long-term memory and general knowledge questions.

We modeled the effects of visual perceptual learning as a simple decrease in sensory noise. There is indeed ample evidence that perceptual learning leads to noise attenuation (Dosher & Lu, 1998, 1999, 2017; Petrov et al., 2005; Raiguel et al., 2006). However, at the same time, perceptual learning may also increase the signal (Solovey, Shalom, Pérez-Schuster, & Sigman, 2016), sharpen the perceptual template used to process the stimulus (Li, Levi, & Klein, 2004), improve probabilistic inference (Bejjanki, Beck, Lu, & Pouget, 2011), and so forth (for reviews, see Dosher & Lu, 2017; Lu, Hua, Huang, Zhou, & Dosher, 2011; Watanabe & Sasaki, 2015). Perceptual learning likely has many consequences and our experiments were not designed to distinguish or weight the importance of each of these effects. Rather, perceptual learning was used as a tool that allowed us to decrease sensory noise in our model. Several previous studies have combined confidence ratings and perceptual learning (Guggenmos, Wilbertz, Hebart, & Sterzer, 2016; Schwiedrzik et al., 2011; Solovey et al., 2016; Zizlsperger, Kümmel, & Haarmeier, 2016), but although they found important effects of learning on the overall confidence level, none investigated how training affects metacognitive efficiency.

An important question for future research is whether metacognitive efficiency can be trained. Given that subjects completed the same metacognitive task for 7 days, one may expect that their metacognitive ability would increase. Our design did not allow us to separate the effects of training on sensory and metacognitive noise, but given the decrease of metacognitive efficiency, putative decreases in metacognitive noise must have been relatively small. This conclusion was further reinforced by the fact that we did not see a change in metacognitive efficiency for the two untrained conditions in Experiment 1. Importantly, we did not include trial-to-trial feedback, and it is perhaps this type of feedback that could allow subjects to improve their metacognitive judgments as shown previously (Maniscalco et al., 2016). Nevertheless, it is possible that metacognitive noise decreased even in our Experiment 1 but that its effect on M_{ratio} was masked by the larger decrease in sensory noise. To isolate the effect of metacognitive noise, future training experiments should include shorter training sessions (or mix in different sensory stimuli) in order to minimize the decrease in sensory noise.

In conclusion, we illustrated the existence of a robust positive relationship between the level of sensory noise and metacognitive efficiency. These results point to the existence of independent metacognitive noise and have strong implications about the meaning and interpretation of metacognitive efficiency.

Context of the Research

This research is part of Rahnev lab's overall research program of elucidating the underlying mechanisms of perceptual decision making. The current project was borne out of our attempts to understand why confidence ratings typically carry less information than the primary decision. This effect is most straightforwardly described as second-level noise, and this research directly tested the existence of such noise. Future research will attempt to specify more precisely the exact nature of this metacognitive noise and determine why it occurs.

References

Bejjanki, V. R., Beck, J. M., Lu, Z.-L. L., & Pouget, A. (2011). Perceptual learning as improved probabilistic inference in early sensory areas. *Nature Neuroscience, 14,* 642–648. http://dx.doi.org/10.1038/nn.2796

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433–436. http://dx.doi.org/10.1163/156856897X00357

David, A. S., Bedford, N., Wiffen, B., & Gilleen, J. (2012). Failures of metacognition and lack of insight in neuropsychiatric disorders. *Philosophical Transactions of the Royal Society of London: Series B, Biological Sciences, 367,* 1379–1390. http://dx.doi.org/10.1098/rstb.2012.0002

Del Cul, A., Dehaene, S., Reyes, P., Bravo, E., & Slachevsky, A. (2009). Causal role of prefrontal cortex in the threshold for access to consciousness. *Brain: A Journal of Neurology, 132,* 2531–2540. http://dx.doi.org/10.1093/brain/awp111

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods, 47,* 1–12. http://dx.doi.org/10.3758/s13428-014-0458-y

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience, 16,* 105–110. http://dx.doi.org/10.1038/nn.3279

Dosher, B. A., & Lu, Z.-L. (1998). Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. *Proceedings of the National Academy of Sciences of the United States of America, 95,* 13988–13993. http://dx.doi.org/10.1073/pnas.95.23.13988

Dosher, B. A., & Lu, Z.-L. (1999). Mechanisms of perceptual learning. *Vision Research, 39,* 3197–3221. http://dx.doi.org/10.1016/S0042-6989(99)00059-0

Dosher, B. A., & Lu, Z.-L. (2017). Visual perceptual learning and models. *Annual Review of Vision Science, 3,* 9.1–9.21. http://dx.doi.org/10.1146/annurev-vision-102016-061249

Fetsch, C. R., Kiani, R., Newsome, W. T., & Shadlen, M. N. (2014). Effects of cortical microstimulation on confidence in a perceptual decision. *Neuron, 83,* 797–804. http://dx.doi.org/10.1016/j.neuron.2014.07.011

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review, 124,* 91–114. http://dx.doi.org/10.1037/rev0000045

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

Fleming, S. M., Ryu, J., Golfinos, J. G., & Blackmon, K. E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain: A Journal of Neurology, 137,* 2811–2822. http://dx.doi.org/10.1093/brain/awu221

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science, 329,* 1541–1543. http://dx.doi.org/10.1126/science.1191883

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

Guggenmos, M., Wilbertz, G., Hebart, M. N., & Sterzer, P. (2016). Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *eLife, 5,* e13388. http://dx.doi.org/10.7554/eLife.13388

Haefner, R. M., Berkes, P., & Fiser, J. (2016). Perceptual decision-making as probabilistic inference by neural sampling. *Neuron, 90,* 649–660. http://dx.doi.org/10.1016/j.neuron.2016.03.020

Hangya, B., Sanders, J. I., & Kepecs, A. (2016). A mathematical framework for statistical decision confidence. *Neural Computation, 28,* 1840–1858. http://dx.doi.org/10.1162/NECO_a_00864

Hauser, T. U., Allen, M., Purg, N., Moutoussis, M., Rees, G., & Dolan, R. J. (2017). Noradrenaline blockade specifically enhances metacognitive performance. *eLife, 6,* e24901. http://dx.doi.org/10.7554/eLife.24901

Higham, P. A., Perfect, T. J., & Bruno, D. (2009). Investigating strength and frequency effects in recognition memory using type-2 signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 57–80. http://dx.doi.org/10.1037/a0013865

Jachs, B., Blanco, M. J., Grantham-Hill, S., & Soto, D. (2015). On the independence of visual awareness and metacognition: A signal detection theoretic analysis. *Journal of Experimental Psychology: Human Perception and Performance, 41,* 269–276. http://dx.doi.org/10.1037/xhp0000026

Jang, Y., Wallsten, T. S., & Huber, D. E. (2012). A stochastic detection and retrieval model for the study of metacognition. *Psychological Review, 119,* 186–200. http://dx.doi.org/10.1037/a0025960

Jolij, J., & Lamme, V. A. F. (2005). Repression of unconscious information by conscious processing: Evidence from affective blindsight induced by transcranial magnetic stimulation. *Proceedings of the National Academy of Sciences of the United States of America, 102,* 10747–10751. http://dx.doi.org/10.1073/pnas.0500834102

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

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science, 220,* 671–680. http://dx.doi.org/10.1126/science.220.4598.671

Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition: An International Journal, 10,* 294–340. http://dx.doi.org/10.1006/ccog.2000.0494

Li, R. W., Levi, D. M., & Klein, S. A. (2004). Perceptual learning improves efficiency by re-tuning the decision 'template' for position discrimination. *Nature Neuroscience, 7,* 178–183. http://dx.doi.org/10.1038/nn1183

Lim, K., Wang, W., & Merfeld, D. M. (2017). Unbounded evidence accumulation characterizes subjective visual vertical forced-choice perceptual choice and confidence. *Journal of Neurophysiology, 118,* 2636–2653. http://dx.doi.org/10.1152/jn.00318.2017

Lu, Z.-L., Hua, T., Huang, C. B., Zhou, Y., & Dosher, B. A. (2011). Visual perceptual learning. *Neurobiology of Learning and Memory, 95,* 145–151. http://dx.doi.org/10.1016/j.nlm.2010.09.010

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

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

Mamassian, P. (2016). Visual confidence. *Annual Review of Vision Science, 2,* 459–481. http://dx.doi.org/10.1146/annurev-vision-111815-114630

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

Maniscalco, B., & Lau, H. (2014). Signal detection theory analysis of Type 1 and Type 2 data: Meta-d′, response-specific meta-d′, and the unequal variance SDT model. In S. M. Fleming & C. D. Frith (Eds.), *The cognitive neuroscience of metacognition* (pp. 25–66). Berlin, Germany: Springer. http://dx.doi.org/10.1007/978-3-642-45190-4_3

Maniscalco, B., & Lau, H. (2015). Manipulation of working memory contents selectively impairs metacognitive sensitivity in a concurrent visual discrimination task. *Neuroscience of Consciousness, 2015,* niv002. http://dx.doi.org/10.1093/nc/niv002

Maniscalco, B., & Lau, H. (2016). The signal processing architecture underlying subjective reports of sensory awareness. *Neuroscience of Consciousness, 2016*(1), niw002. http://dx.doi.org/10.1093/nc/niw002

Maniscalco, B., McCurdy, L. Y., Odegaard, B., & Lau, H. (2017). Limited cognitive resources explain a trade-off between perceptual and metacognitive vigilance. *The Journal of Neuroscience, 37,* 1213–1224. http://dx.doi.org/10.1523/JNEUROSCI.2271-13.2016

Maniscalco, B., Peters, M. A. K., & Lau, H. (2016). Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Attention, Perception, & Psychophysics, 78,* 923–937. http://dx.doi.org/10.3758/s13414-016-1059-x

Metcalfe, J., & Shimamura, A. P. (1994). *Metacognition: Knowing about Knowing*. Cambridge, MA: MIT Press.

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review, 15,* 465–494. http://dx.doi.org/10.3758/PBR.15.3.465

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95,* 109–133. http://dx.doi.org/10.1037/0033-2909.95.1.109

Orbán, G., Berkes, P., Fiser, J., & Lengyel, M. (2016). Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron, 92,* 530–543. http://dx.doi.org/10.1016/j.neuron.2016.09.038

Persaud, N., Davidson, M., Maniscalco, B., Mobbs, D., Passingham, R. E., Cowey, A., & Lau, H. (2011). Awareness-related activity in prefrontal and parietal cortices in blindsight reflects more than superior visual performance. *NeuroImage, 58,* 605–611. http://dx.doi.org/10.1016/j.neuroimage.2011.06.081

Peters, M. A. K., Thesen, T., Ko, Y. D., Maniscalco, B., Carlson, C., Davidson, M., . . . Lau, H. (2017). Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature Human Behaviour, 1,* 0139. http://dx.doi.org/10.1038/s41562-017-0139

Petrov, A. A., Dosher, B. A., & Lu, Z.-L. L. (2005). The dynamics of perceptual learning: An incremental reweighting model. *Psychological Review, 112,* 715–743. http://dx.doi.org/10.1037/0033-295X.112.4.715

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

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience, 19,* 366–374. http://dx.doi.org/10.1038/nn.4240

Qamar, A. T., Cotton, R. J., George, R. G., Beck, J. M., Prezhdo, E., Laudano, A., . . . Ma, W. J. (2013). Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. *Proceedings of the National Academy of Sciences of the United States of America, 110,* 20332–20337. http://dx.doi.org/10.1073/pnas.1219756110

Rahnev, D. A., Bahdo, L., de Lange, F. P., & Lau, H. (2012). Prestimulus hemodynamic activity in dorsal attention network is negatively associated with decision confidence in visual perception. *Journal of Neurophysiology, 108,* 1529–1536. http://dx.doi.org/10.1152/jn.00184.2012

Rahnev, D., & Denison, R. N. (2018). Suboptimality in perceptual decision making. *Behavioral and Brain Sciences, 2018,* 1–107. http://dx.doi.org/10.1017/S0140525X18000936

Rahnev, D., Koizumi, A., McCurdy, L. Y., D'Esposito, M., & Lau, H. (2015). Confidence leak in perceptual decision making. *Psychological Science, 26,* 1664–1680. http://dx.doi.org/10.1177/0956797615595037

Rahnev, D., Kok, P., Munneke, M., Bahdo, L., de Lange, F. P., & Lau, H. (2013). Continuous theta burst transcranial magnetic stimulation reduces resting state connectivity between visual areas. *Journal of Neurophysiology, 110,* 1811–1821. http://dx.doi.org/10.1152/jn.00209.2013

Rahnev, D., Maniscalco, B., Graves, T., Huang, E., de Lange, F. P., & Lau, H. (2011). Attention induces conservative subjective biases in visual perception. *Nature Neuroscience, 14,* 1513–1515. http://dx.doi.org/10.1038/nn.2948

Rahnev, D. A., Maniscalco, B., Luber, B., Lau, H., & Lisanby, S. H. (2012). Direct injection of noise to the visual cortex decreases accuracy but increases decision confidence. *Journal of Neurophysiology, 107,* 1556–1563. http://dx.doi.org/10.1152/jn.00985.2011

Rahnev, D., Nee, D. E., Riddle, J., Larson, A. S., & D'Esposito, M. (2016). Causal evidence for frontal cortex organization for perceptual decision making. *Proceedings of the National Academy of Sciences of the United States of America, 113,* 6059–6064. http://dx.doi.org/10.1073/pnas.1522551113

Raiguel, S., Vogels, R., Mysore, S. G., & Orban, G. A. (2006). Learning to see the difference specifically alters the most informative V4 neurons. *The Journal of Neuroscience, 26,* 6589–6602. http://dx.doi.org/10.1523/JNEUROSCI.0457-06.2006

Rounis, E., Maniscalco, B., Rothwell, J. C., Passingham, R. E., & Lau, H. (2010). Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive Neuroscience, 1,* 165–175. http://dx.doi.org/10.1080/17588921003632529

Ryals, A. J., Rogers, L. M., Gross, E. Z., Polnaszek, K. L., & Voss, J. L. (2016). Associative recognition memory awareness improved by theta-burst stimulation of frontopolar cortex. *Cerebral Cortex, 26,* 1200–1210. http://dx.doi.org/10.1093/cercor/bhu311

Sanders, J. I., Hangya, B., & Kepecs, A. (2016). Signatures of a statistical computation in the human sense of confidence. *Neuron, 90,* 499–506. http://dx.doi.org/10.1016/j.neuron.2016.03.025

Schwiedrzik, C. M., Singer, W., & Melloni, L. (2011). Subjective and objective learning effects dissociate in space and in time. *Proceedings of the National Academy of Sciences of the United States of America, 108,* 4506–4511. http://dx.doi.org/10.1073/pnas.1009147108

Seitz, A. R., Kim, D., & Watanabe, T. (2009). Rewards evoke learning of unconsciously processed visual stimuli in adult humans. *Neuron, 61,* 700–707. http://dx.doi.org/10.1016/j.neuron.2009.01.016

Semmelmann, K., & Weigelt, S. (2017). Online psychophysics: Reaction time effects in cognitive experiments. *Behavior Research Methods, 49,* 1241–1260. http://dx.doi.org/10.3758/s13428-016-0783-4

Shekhar, M., & Rahnev, D. (2018). Distinguishing the roles of dorsolateral and anterior PFC in visual metacognition. *The Journal of Neuroscience, 38,* 5078–5087. http://dx.doi.org/10.1523/JNEUROSCI.3484-17.2018

Shibata, K., Sasaki, Y., Bang, J. W., Walsh, E. G., Machizawa, M. G., Tamaki, M., . . . Watanabe, T. (2017). Overlearning hyperstabilizes a skill by rapidly making neurochemical processing inhibitory-dominant. *Nature Neuroscience, 20,* 420–425. http://dx.doi.org/10.1038/nn.4490

Shibata, K., Watanabe, T., Sasaki, Y., & Kawato, M. (2011). Perceptual learning incepted by decoded fMRI neurofeedback without stimulus presentation. *Science, 334,* 1413–1415. http://dx.doi.org/10.1126/science.1212003

Shimamura, A. P. (2000). Toward a cognitive neuroscience of metacognition. *Consciousness and Cognition, 9,* 313–323. http://dx.doi.org/10.1006/ccog.2000.0450

Solovey, G., Shalom, D., Pérez-Schuster, V., & Sigman, M. (2016). Perceptual learning effect on decision and confidence thresholds. *Consciousness and Cognition: An International Journal, 45,* 24–36. http://dx.doi.org/10.1016/j.concog.2016.08.010

van den Berg, R., Yoo, A. H., & Ma, W. J. (2017). Fechner's law in metacognition: A quantitative model of visual working memory confidence. *Psychological Review, 124,* 197–214. http://dx.doi.org/10.1037/rev0000060

Watanabe, T., & Sasaki, Y. (2015). Perceptual learning: Toward a comprehensive theory. *Annual Review of Psychology, 66,* 197–221. http://dx.doi.org/10.1146/annurev-psych-010814-015214

Weiskrantz, L. (1996). Blindsight revisited. *Current Opinion in Neurobiology, 6,* 215–220. http://dx.doi.org/10.1016/S0959-4388(96)80075-4

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society of London: Series B, Biological Sciences, 367,* 1310–1321. http://dx.doi.org/10.1098/rstb.2011.0416

Zizlsperger, L., Kümmel, F., & Haarmeier, T. (2016). Metacognitive confidence increases with, but does not determine, visual perceptual learning. *PLoS ONE, 11*(3), e0151218. http://dx.doi.org/10.1371/journal.pone.0151218

Received December 14, 2017
Revision received August 16, 2018
Accepted August 17, 2018 ■