*Research Article*

Psychological Science
2015, Vol. 26(11) 1664–1680
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797615595037
pss.sagepub.com

# Confidence Leak in Perceptual Decision Making

**Dobromir Rahnev[1,2], Ai Koizumi[3], Li Yan McCurdy[4], Mark D'Esposito[2], and Hakwan Lau[5,6]**

[1]Department of Psychology, Georgia Institute of Technology; [2]Helen Wills Neuroscience Institute, University of California, Berkeley; [3]Department of Psychology, Columbia University; [4]Interdepartmental Neuroscience Program, Yale University School of Medicine; [5]Department of Psychology, University of California, Los Angeles; and [6]Brain Research Institute, University of California, Los Angeles

## Abstract

People live in a continuous environment in which the visual scene changes on a slow timescale. It has been shown that to exploit such environmental stability, the brain creates a *continuity field* in which objects seen seconds ago influence the perception of current objects. What is unknown is whether a similar mechanism exists at the level of metacognitive representations. In three experiments, we demonstrated a robust intertask *confidence leak*—that is, confidence in one's response on a given task or trial influencing confidence on the following task or trial. This confidence leak could not be explained by response priming or attentional fluctuations. Better ability to modulate confidence leak predicted higher capacity for metacognition as well as greater gray matter volume in the prefrontal cortex. A model based on normative principles from Bayesian inference explained the results by postulating that observers subjectively estimate the perceptual signal strength in a stable environment. These results point to the existence of a novel metacognitive mechanism mediated by regions in the prefrontal cortex.

## Keywords

perception, decision making, attention, open data, open materials

Received 10/15/14; Revision accepted 6/15/15

For various kinds of decisions, people have the ability not only to use the available information to make a choice between several alternatives, but also to consider the likelihood that their decision is correct. This metacognitive ability is critical in deciding whether to act immediately on the decision or to continue gathering information, as well as whether to update one's model of the world with the newly acquired information (Fleming, Dolan, & Frith, 2012; Koriat, 2007; Metcalfe & Shimamura, 1994; Nelson & Narens, 1990; Shimamura, 2000; Yeung & Summerfield, 2012). Research on metacognitive judgments in perception has typically focused on how confidence ratings are influenced by different properties of the stimulus at hand (Fleming & Lau, 2014). However, other factors beyond the immediate stimulus also influence confidence ratings (Koriat, 2011). One such factor that has received little attention is intertrial and intertask influences on metacognitive judgments.

People live in a continuous environment in which viewing conditions change relatively slowly and objects tend to persist within the visual field for long periods. It has been demonstrated that the brain exploits such environmental stability (Fischer & Whitney, 2014; Fründ, Wichmann, & Macke, 2014; Liberman, Fischer, & Whitney, 2014; Zhang, Wang, & Goldberg, 2014). In particular, it has recently been proposed that the brain creates a *continuity field* in which recently seen objects bias the perception of current objects (Fischer & Whitney, 2014; Liberman et al., 2014).

Here, we report evidence that such continuity fields exist at the metacognitive level and that there may be dedicated mechanisms specifically influencing confidence ratings rather than basic perceptual processing. We demonstrated robust intertask dependence in metacognitive judgments of confidence even when the

**Corresponding Author:**
Dobromir Rahnev, Georgia Institute of Technology, 130 Coon Building, 654 Cherry St., Atlanta, GA 30332
E-mail: drahnev@gmail.com

Downloaded from pss.sagepub.com at GEORGIA TECH LIBRARY on February 12, 2016

two different tasks involve distinct visual features: letter identity and color. In other experiments, we found that serial dependence of confidence ratings appeared when the same task was presented many times. Finally, we related this effect to gray matter volume in the anterior prefrontal cortex (aPFC), an area that has previously been linked to metacognition. This phenomenon—which we call *confidence leak*—is explained by a computational model based on the principles of Bayesian inference. The model quantifies the normative principle that observers use their subjective certainty in previous judgments to predict the quality of the perceptual signal for future judgments.

## Method

### Observers

Sixty-nine observers (39 women, 30 men; mean age = 23.6 years, *SD* = 5.7) participated in three psychophysical experiments (Experiments 1–3). Twenty-seven observers completed Experiment 1, but 1 observer was excluded for chance performance. Twenty-two observers completed Experiment 2, but 4 observers were excluded for having an extreme bias in the opt-out task: 2 of them chose the opt-out option on at least 396 trials (out of 400), whereas the other 2 chose to opt out at most 5 times. Experiment 3 was a reanalysis of a 20-observer data set previously reported in Rahnev, Maniscalco, et al. (2011). For all three studies, we sought to collect data from between 20 and 30 observers, as in previous work from our laboratory (Maniscalco, Bang, Iravani, Camps-Febrer, & Lau, 2012; Rahnev, Lau, & De Lange, 2011). Data collection stopped when the sample size reached the target interval; logistical factors led to slightly different sample sizes. No statistical analyses reported in this article were performed on partial data. In addition, in Experiment 4, we reanalyzed the data from McCurdy et al. (2013), which included 34 observers. All were naive to the purpose of the experiments, had normal or corrected-to-normal vision, and signed an informed-consent statement approved by the local ethics committee. Observers were compensated at the rate of $10 per hr.

### Procedure

Stimuli in Experiments 1 through 3 were presented on a gray background (6.0 cd/m$^2$). Observers were seated in a dim room about 60 cm away from the computer monitor. Stimuli were generated using the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) in MATLAB (The MathWorks, Natick, MA) and were shown on a 19-in. iMac monitor (1,680 × 1,050 pixel resolution, 60-Hz refresh rate). Each experiment took approximately 1 hr to complete.

***Experiment 1.*** On each trial in Experiment 1, 40 characters were displayed simultaneously in a random pattern. The characters were *X*s and *O*s colored red and blue. Letter identity and color were independent of each other. Observers' task was to judge the dominant letter identity and color in the display and to indicate their confidence in those judgments (Fig. 1). More specifically, observers answered four questions on each trial: (a) whether there were more *X*s or more *O*s, (b) how confident they were in their letter-identity decision on a scale from 1 to 4, (c) whether there were more red or blue characters, and (d) how confident they were in their color decision on a scale from 1 to 4. The order of the four questions was the same for all observers. To give their responses, observers used the 1 through 4 keys on a computer keyboard.

The dominant letter always accounted for 23 of the 40 characters on the screen. In different runs, the dominant color accounted for either 23 or 29 characters. The dominant letter and color were pseudorandomized, but over the course of the experiment, the *X*s and *O*s, as well as the red and blue color, were dominant equally often. We used the same objective difficulty level for all observers because achieving very similar levels of performance across observers was not critical. Conversely, the use of a staircase procedure could have led observers to believe that a correlation structure was actually present in the task (even though such structure did not exist in the main experiment), thus potentially compromising our analyses on confidence leak.

All letters were presented in Arial font (size = 0.5°) and placed randomly within an imaginary square centered on fixation; each side of the square subtended 10° of visual angle. The letters remained on the screen for 1 s, after which each of the four questions was presented on the screen in succession. Observers were allowed to take as long as they needed to give their responses.

Observers completed a total of 400 trials across four runs, each run consisting of four blocks of 25 trials each. At the end of each block, observers were given a 15-s break, while at the end of each run, they were allowed to take self-paced breaks. Observers were given a total of 46 practice trials. The initial practice trials were easier than the later ones, and the difficulty was gradually increased until it reached the difficulty of the actual experiment. Trial-by-trial feedback was provided on the first 36 trials to help observers learn how to perform the task. There was no feedback in the last 10 practice trials or in the main experiment.

***Experiment 2.*** Experiment 2 was similar to Experiment 1. The main difference was the method of collecting confidence ratings. In Experiment 1, confidence ratings were always collected on the same scale from 1 to

Downloaded from pss.sagepub.com at GEORGIA TECH LIBRARY on February 12, 2016

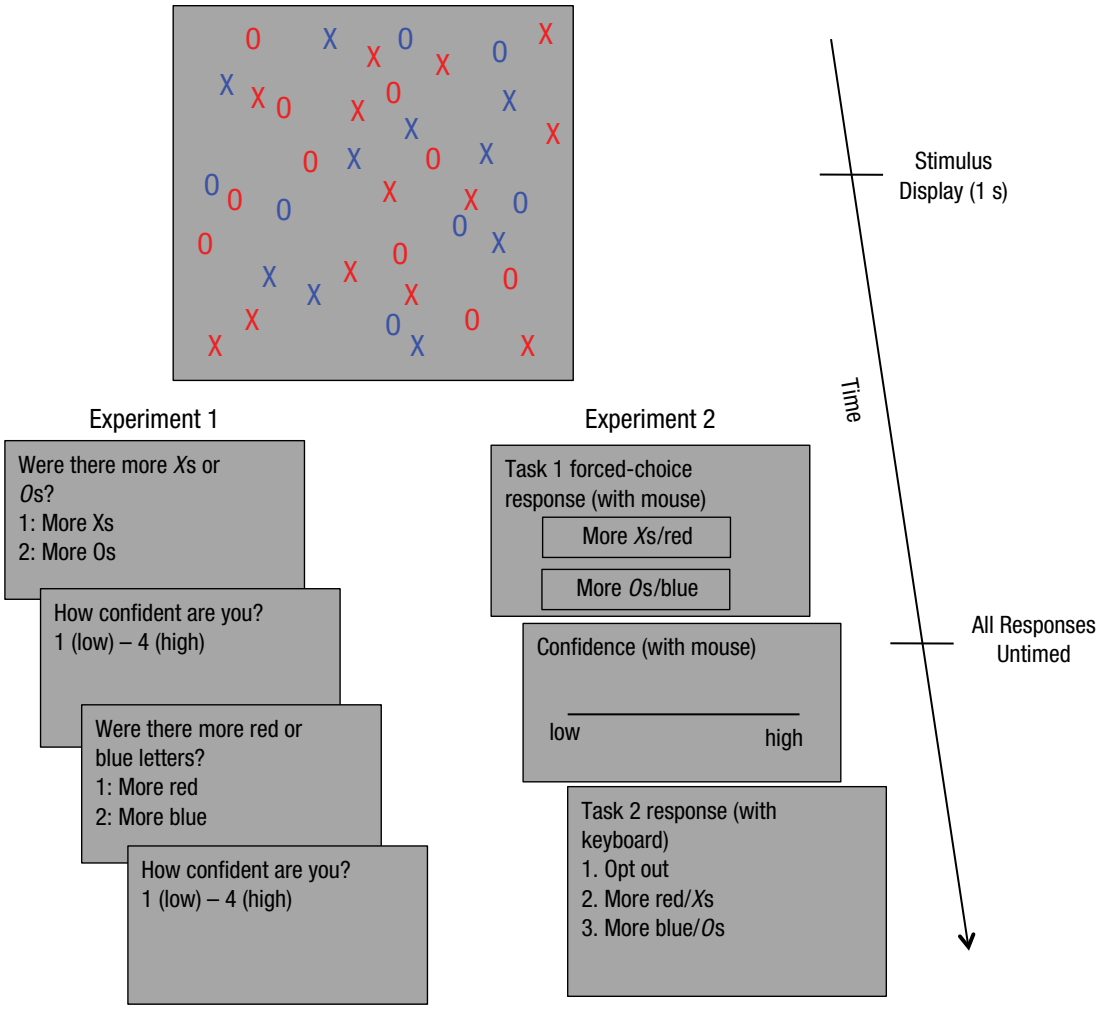


**Fig. 1.** Tasks for Experiments 1 and 2. The stimulus consisted of 40 *X*s and *O*s colored red and blue. On each trial, observers had to decide whether there were more *X*s or more *O*s (letter-identity task), as well as whether there were more red or blue letters (color task). Observers also indicated their level of confidence after each task. In Experiment 1, confidence on both tasks was rated separately on a scale from 1 to 4. In Experiment 2, for one of the tasks, observers used a visual analog scale (VAS) by sliding a marker; for the other task, they decided whether to provide an answer in order to win a larger reward (thereby indicating a high level of confidence) or to opt out and not give a response, thus earning a smaller, guaranteed reward (thereby indicating a low level of confidence). To further minimize response priming in Experiment 2, we had observers make the VAS response with a mouse and the opt-out response with a keyboard.

4. In Experiment 2, in order to minimize the chance of motor priming between confidence reporting in the two different tasks, we used two separate methods of determining the confidence level. These methods were designed to be as different as possible from each other. The first reporting method required observers to use the mouse to make a subjective confidence response on a continuous visual analog scale (VAS), while the second was the opt-out paradigm (Kiani & Shadlen, 2009), in which observers were given the option of not responding if they were unsure about the response (but did not give confidence explicitly).

Observers gave their confidence rating on the VAS by moving a mouse between the two extremes of a straight line. The left extreme was marked "not confident at all" and was coded as a confidence of 0, while the right extreme was marked "very confident" and was coded as a confidence of 100. For each observer, this VAS rating consistently followed either the letter-identity or color task, but this pairing was randomized between observers. On the other task, observers used the keyboard to select one of three choices: They could indicate whether there were more red letters (or more *X*s), whether there were more blue letters (or more *O*s), or whether they wanted to opt out of responding. The presentation order of the letter-identity and color tasks was counterbalanced. The decision to opt out was coded as a confidence level of 1, while giving a response was coded as a confidence level of 2. In order to make the opt-out choice meaningful, we introduced a point system. Choosing to opt out resulted in 2

Downloaded from pss.sagepub.com at GEORGIA TECH LIBRARY on February 12, 2016

points guaranteed, while choosing to respond led to either 4 points (for correct answers) or −1 points (for incorrect answers). The optimal strategy in this task was thus to choose to opt out when the probability of being correct was less than 66%. To decrease interobserver variability, we explicitly informed observers of the optimal strategy. To increase the consistency between the two tasks, we awarded 4 points for correct and −1 points for incorrect answers in the task that was followed by the VAS rating. To motivate observers to use the opt-out option optimally, we rewarded the 3 observers with highest scores across the whole experiment with an additional $10.

To remove any other biases potentially present in Experiment 1, we randomized the order of the questions between observers such that about half of them would always respond first to the letter-identity task, while the other half responded first to the color task. Further, in Experiment 2, both tasks had two difficulty levels (the dominant letter identity or color was present in either 23 or 26 of the 40 characters), which were presented in completely random order across trials.

***Experiment 3.*** Experiment 3 was performed to investigate whether confidence leak depends on the quality of the perceptual signal. It was originally reported in the Supplemental Material in Rahnev, Maniscalco, et al. (2011). There, we focused on comparing the difference in confidence between high-attention (two stimuli) and low-attention (four stimuli) conditions. We returned to this data set to reanalyze it in terms of confidence leak.

In that experiment, we varied the number of stimuli (Gabor patches) on the screen to manipulate how observers distribute their attention to different objects. In one condition, we used two items on the screen (a relatively focused mode of attention), while in the other we used four items on the screen (a relatively distributed mode of attention). The stimuli were presented for 33 ms (two computer frames). After a delay of 500 ms, observers saw a response cue that instructed them which stimulus they should respond to. Observers had to indicate the tilt (clockwise/counterclockwise) of the Gabor patch and rate their confidence (high/low). Observers completed eight blocks of 125 trials each for a total of 1,000 trials. Within each block, there were always either two or four patches and a single contrast level (4, 6, 8, 10, or 12%).

***Experiment 4.*** Experiment 4 was a reanalysis of the data from McCurdy et al. (2013). All experimental details are included in the original publication. Very briefly, observers completed a two-alternative forced-choice task in which they indicated which of two noisy stimuli located on the left and right of fixation contained a grating. They then provided a confidence rating on a scale from 1 to 4. Observers completed 510 trials separated in five blocks of 102 trials each.

## Analyses

***Experiment 1.*** To determine whether confidence leaks from one task to the other, we first performed, for each observer, a simple trial-by-trial correlation of the confidence ratings from the two tasks in Experiment 1. Next, to control for the influence of other factors, we performed a regression in which confidence on the letter-identity task was used to predict confidence on the color task while at the same time controlling for the influence of accuracy and reaction time (RT) on each task, as well as difficulty of the color task (i.e., whether the dominant color was present in 23 or 29 characters).

One possible reason for a confidence leak is simple motoric priming. To control for that possibility, we analyzed all trials in which an observer gave a confidence rating of 1 in one of the tasks but not 1 in the other task. This allowed us to determine, for each observer, the percentage of times that confidence ratings of 2, 3, and 4 on one task were paired with a confidence rating of 1 on the other task. Similarly, we determined the percentage of times that confidence ratings of 1, 2, and 3 on one task were paired with a confidence rating of 4 on the other task. We excluded 6 observers because they did not have at least five trials in either of these analyses, and including them would have led to excessively volatile estimates. Finally, we performed a repeated measures analysis of variance (ANOVA) to test whether there was an interaction between the number of 2, 3, and 4 confidence ratings paired with a confidence of 1 and the number of 1, 2, and 3 confidence ratings paired with a confidence of 4.

***Experiment 2.*** For Experiment 2, we performed correlation and regression analyses as in Experiment 1. In addition, for visualization purposes, we created probability density functions (PDFs) for confidence on the VAS conditional on confidence on the opt-out task. The individual PDFs were averaged and, for visualization, smoothed with a 10-point window.

***Experiments 3 and 4.*** Because observers completed only one task per trial, confidence leak in Experiments 3 and 4 was determined by analyzing the temporal Lag-1 autocorrelation of the confidence ratings. As in Experiments 1 and 2, both correlation and regression analyses were performed. To determine observers' performance on the task, we computed the $d'$ measure from signal detection theory by calculating the hit rate (HR) and false alarm rate (FAR). These data were entered into Equation 1,

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

where $\phi^{-1}$ is the inverse of the cumulative standard normal distribution that transforms HR and FAR into $z$ scores.

Downloaded from pss.sagepub.com at GEORGIA TECH LIBRARY on February 12, 2016

The measure $d'$ reflects the signal-to-noise ratio for observers performing the task.

***Bayes factors.*** We computed Bayes factors for confidence leak in each experiment. We employed the Bayes calculator in Dienes (2008), for which we used a wide prior distribution for the alternative hypothesis defined as a half-normal distribution with a mean of 0 and a standard deviation of 1. Values higher than 3 are usually regarded as substantial evidence for the alternative hypothesis (Dienes, 2008, 2014). We obtained very large Bayes factors, and the results were insensitive to variations in the distribution used for the alternative hypothesis.

***Metacognition.*** To understand the relationship between confidence leak and metacognition, we computed a standard measure for metacognition, the area under the Type 2 receiver-operating-characteristic (ROC) curve (Fleming & Lau, 2014; Fleming, Weil, Nagy, Dolan, & Rees, 2010). The Type 2 ROC curve is similar to a conventional ROC curve, with the difference that hits are defined as trials on which responses are correct and confidence is high, while false alarms are defined as trials on which responses are incorrect and confidence is high. The area under the Type 2 ROC curve (Type 2 AUC) was computed for each of the two tasks in Experiment 1 (the average of the two was taken as the observer-specific metacognition score; similar results were obtained when both scores were considered separately). In Experiment 2, the Type 2 AUC score was computed only for the task on which observers rated their confidence, because the opt-out task did not allow us to determine the accuracy of the low-confidence decisions (a decision to opt out meant that observers did not indicate the perceived stimulus). For this analysis, VAS confidence scores were transformed into scores from 1 to 4 using cutoffs defined on the 25th, 50th, and 75th percentile of VAS scores for each individual observer. Finally, in Experiment 3, Type 2 AUC was computed using all trials from the experiment.

We then correlated, across all observers from Experiments 1 through 3, the amount of confidence leak, defined as the Fisher-transformed correlation values from each experiment, with the observer-specific metacognition (defined as the Type 2 AUC value). This correlation was performed on the combined data from the three experiments to increase power.

***Voxel-based morphometry.*** We repeated the analyses by McCurdy et al. (2013), except that we were interested in the brain correlates of confidence leak rather than in metacognition scores. Briefly, we used the same preprocessing procedure, which included segmenting the data into gray matter, white matter, and cerebrospinal fluid in native space; aligning the data from different observers; registering the coordinates to Montreal Neurological Institute stereotactic space; and smoothing the data with an 8-mm full-width at half-maximum Gaussian kernel. Multiple regression was used, and an initial threshold of $p < .001$, uncorrected, was set to determine the brain regions that correlated with our measures of confidence leak. Gender was included as a covariate, and proportional scaling was used to account for variability in global brain volume across observers. Following McCurdy et al. (2013), we applied small-volume correction to the clusters of interest in the prefrontal cortex by defining a 10-mm sphere at the peak voxel coordinates that Fleming et al. (2010) found to be associated with their measure of metacognitive capacity—left aPFC: $x = -20$, $y = 53$, $z = 12$; right aPFC: $x = 24$, $y = 65$, $z = 18$; $x = 33$, $y = 50$, $z = 9$; right dorsolateral prefrontal cortex (dorsolateral PFC): $x = 36$, $y = 39$, $z = 21$.

## Modeling framework

To explain the confidence-leak phenomenon, we developed a single-parameter model and fitted it with the data from the first three experiments. The model was based on the principles of Bayesian inference and was an extension of our previous work on confidence generation (Rahnev, Bahdo, de Lange, & Lau, 2012; Rahnev et al., 2013; Rahnev, Maniscalco, et al., 2011; Rahnev, Maniscalco, Luber, Lau, & Lisanby, 2012)

Bayesian theory describes the optimal way to choose between different alternatives in the space of possible stimuli. Observers attempt to make a decision on the basis of the posterior probability $p(S_i|E = x)$, where $S_i$ represents the possible stimulus category, and $E$ is the evidence on the current trial. Bayes' theorem describes how $p(S_i|E = x)$ should be computed based on the likelihood function $f_{E|S_i}(x)$, the prior probability $p(S_i)$, and the density function of $E$, $f_E(x)$:

$$p(S_i|E = x) = \frac{f_{E|S_i}(x) \times p(S_i)}{f_E(x)}. \quad (2)$$

For all perceptual decisions in the current set of experiments, there were two possible stimulus categories, $S_1$ and $S_2$ (in Experiments 1 and 2, these were either $X/O$ or red/blue, while in Experiment 3, these were clockwise/counterclockwise tilt). Because we informed observers that the two stimulus categories were equally likely in all experiments, we set the prior probabilities for the two stimulus categories to .5. Further, because there were only two stimulus categories, $p(S_2 \mid E) = 1 - p(S_1 \mid E)$. Thus, observers' choice was made by simply comparing $p(S_1 \mid E)$ and $p(S_2 \mid E)$ and choosing the stimulus that corresponded to the higher posterior probability. In

Downloaded from pss.sagepub.com at GEORGIA TECH LIBRARY on February 12, 2016

keeping with previous literature (Macmillan & Creelman, 2005), we assumed the likelihood functions $f_{E|S_i}(x)$ to be Gaussian distributions with equal standard deviations that could be set to 1:

$$f_{E|S_i}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2}}, \quad (3)$$

where $\mu_i$ is the mean of the likelihood function $f_{E|S_i}(x)$. For simplicity, and without loss of generality, we set the two means to be symmetric around the origin, that is $\mu_1 = -\mu$, $\mu_2 = \mu$, where $\mu$ is positive. Notice that with these assumptions, $d'$ can be expressed as a function of $\mu$:

$$d' = 2 \times \mu. \quad (4)$$

Using Equation 3, we derived the following equation for the posterior distribution $p(S_2 \mid E = x)$ (for the full derivation, see the Supplemental Material available online):

$$p(S_2 | E = x) = \frac{1}{e^{-2x\mu} + 1}. \quad (5)$$

How are confidence ratings determined in this framework? According to the principles of Bayesian inference, confidence ratings should be determined by the probability of being correct. Thus, for example, if an observer is given the option of giving low or high confidence, then he or she would make this choice based on the maximum value of $p(S_i|E)$, $\max[p(S_i|E)]$, and give a high confidence rating if this posterior probability is higher than a threshold (e.g., .75) and a low confidence rating otherwise. In general, if observers are asked to produce $N$ discrete confidence ratings, they would need to define $N - 1$ different thresholds $t_1, t_2, \ldots t_{N-1}$, such that a confidence rating of $k$ is given if $\max[p(S_i|E)]$ is in the interval $(t_{k-1}, t_k)$, where $t_0 = .5$ and $t_N = 1$ (Macmillan & Creelman, 2005).

The thresholds $t_i$ are defined in the posterior probability space. However, observers can infer the posterior probability space only by assuming a particular shape of the likelihood functions $f_{E|S_i}(x)$. If this assumption is wrong, then the inferred posterior probability space would be incorrect. In this sense, observers do not have direct access to the posterior probability space. Therefore, it is helpful to translate the thresholds $t_i$ from the posterior space to the likelihood space to which observers do have direct access, because that space is independent of any assumptions on the shape of the likelihood functions or the exact priors. Such correspondence is fortunately straightforward. Each threshold $t_i$ corresponds to a criterion $c_i$ defined in likelihood space. Given Equation (5),

$$t_i = \frac{1}{e^{-2c_i\mu} + 1}, \quad (6)$$

from which it follows that

$$c_i = \frac{\log\left(\frac{t_i}{1-t_i}\right)}{2\mu}. \quad (7)$$

The likelihood-space confidence criteria $c_1, c_2, \ldots c_{N-1}$ are placed on the right of the origin in likelihood space; a confidence rating of $k$ is given if evidence mapped on the likelihood scale is in the interval $(c_{k-1}, c_k)$, where $c_0 = 0$ and $c_N = \infty$. These confidence criteria reflect the confidence rating when $S_2$ is the more likely stimulus. In a symmetrical fashion, $-c_1, -c_2, \ldots -c_N$ are used for the confidence ratings when $S_1$ is the more likely stimulus.

With no trial-to-trial information about the likely state of the environment, an observer can simply fix $c_1, c_2, \ldots c_{N-1}$ on the basis of constant prior and likelihood functions so that they correspond to the desired $t_1, t_2, \ldots t_{N-1}$. However, the environment has higher order structure that makes it predictable (David, Vinje, & Gallant, 2004), and the brain is able to take advantage of this predictability (Fischer & Whitney, 2014; Fründ et al., 2014; Liberman et al., 2014; Yu & Cohen, 2009). To account for such autocorrelation in the environment, an observer will need to update the likelihood functions $f_{E|S_i}(x)$ from trial to trial. If observers expect that the difficulty on the subsequent trial is now described by a likelihood function $f'_{E|S_i}(x)$ with a mean of $\mu'$, then they ought to update their criteria $c_i$ to $c'_i$ as follows:

$$c'_i = \frac{\log\left(\frac{t_i}{1-t_i}\right)}{2\mu'} = c_i \times \frac{\mu}{\mu'}. \quad (8)$$

In other words, if observers expect $d'$ to change by a factor of $F$, then to remain Bayes optimal, they need to multiply all criteria, defined in the likelihood space, by a factor of $\frac{1}{F}$ (using Equation 8 and the linear dependency between $d'$ and $\mu$ from Equation 4).

## *Model fitting*

We developed a formal model that is based on the Bayes-optimal strategy of adjusting confidence criteria on the basis of the expected change in $d'$. The model had a single free parameter $\theta$, while all other parameters were fixed. For all experiments, we first computed $d'$ and $c_i$ for each task. The value of $d'$ was computed using Equation 1, whereas $c_i$ was computed using the formula:

Downloaded from pss.sagepub.com at GEORGIA TECH LIBRARY on February 12, 2016

$$c_i = -\frac{1}{2} \times [\Phi^{-1}(\mathrm{HR}_i) + \Phi^{-1}(\mathrm{FAR}_i)], \quad (9)$$

where $\mathrm{HR}_i$ is the proportion of correct responses produced with confidence of $i$ or higher, and $\mathrm{FAR}_i$ is the proportion of false alarms produced with confidence of $i$ or higher (Macmillan & Creelman, 2005).

To perform the model fitting, for each task, we generated random samples from the distributions of the stimulus categories $S_1$ and $S_2$ defined by Equation 3. For each task $k$ ($k$ = 1 or 2, as there were two tasks), we set the means of $S_{1,k}$ and $S_{2,k}$ such that $\mu_{1,k} = -\frac{d'_k}{2}$ and $\mu_{2,k} = \frac{d'_k}{2}$. Because $\mu_{1,k}$ and $\mu_{2,k}$, as well as the criteria $c_{i,k}$, are symmetrical, we could simply sample from the $S_{2,k}$ distributions without loss of generality. The samples for the first task, $x_{j,1}$, were categorized into confidence ratings using the confidence criteria $c_{i,1}$. However, depending on the confidence rating produced on the first task, the criteria for the second task on the corresponding trials are shifted. The idea is that on a trial-by-trial basis, using their confidence in the first task, observers adjust their expectation for the signal-to-noise ratio ($d'$) in the second task as follows:

$$d'_{\text{expected}} = d' \times e^{\Delta C \times \theta}, \quad (10)$$

where $\Delta C$ is deviation from the average observer-specific confidence produced on the same trial in the first task, and $\theta$ is a free parameter that controls the amount of adjustment in the expectation of $d'$ in the second task based on the confidence on the first task. For example, if an observer has an average confidence of 2.5 on the first task, and on a particular trial he or she gave a confidence of 4 on that task, then $\Delta C = 4 - 2.5 = 1.5$. If, for the same observer, $\theta = 0.2$, then $e^{\Delta C \times \theta} = e^{1.5 \times 0.2} = 1.35$. In other words, on this particular trial, $d'_{\text{expected}}$ for the second task would be 1.35 times higher than its normal value. Note that $\theta = 0$ indicates no expectation for any change ($d'_{\text{expected}} = d'$), whereas $\theta > 0$ and $\theta < 0$ indicate expectation for an autocorrelated and antiautocorrelated environment, respectively.

Using Equations 4 and 8, it follows that given the expectation for a change in $d'$ from Equation 10, the observer should update his or her confidence criteria on the second task as follows:

$$c_{i\_\text{new}} = \frac{c_{i\_\text{original}}}{e^{\Delta C \times \theta}}. \quad (11)$$

For each observer in each experiment, the single free parameter $\theta$ was fitted by starting with a value of 0 and adjusting it in a stepwise fashion until we could reproduce the confidence correlation between the two tasks. For each step, we simulated 100,000 trials using the above procedure and adjusted the parameter $\theta$ based on the obtained confidence correlation: If the simulated correlation was smaller or larger than the observed correlation, $\theta$ was initially increased or decreased by a step of 0.2, and this step was decreased in half after each reversal until we obtained a fit of the confidence correlation value with error smaller than 0.001.

In Experiment 1, we computed, for each observer, $d'$ and the confidence criteria $c_1$, $c_2$, and $c_3$ separately for both the letter-identity and color task. The trial-to-trial confidence responses on the letter-identity task (first task) were used to adjust the confidence criteria on the color task (second task).

In Experiment 2, we computed, for each observer, $d'$ and the confidence criteria $c_1$, $c_2$, . . . $c_{100}$ for the VAS so that confidence ratings from 0 to 100 could be produced. For the opt-out task, we could not estimate $d'$ for each observer for reasons outlined earlier, so we set $d'$ for the opt-out task to the same value for each observer that was equal to the average $d'$ for VAS. However, we did compute the single confidence criterion $c$ for each observer separately. The trial-to-trial confidence responses on the opt-out task (first task) were used to adjust the confidence criteria on the VAS task (second task).

In Experiment 3, we computed, for each observer, $d'$ and $c$ separately for the two-stimulus and four-stimulus tasks. On each step, we generated 50,000 trials for each of the two- and four-stimulus tasks (for a total of 100,000 trials). For both the two- and four-stimulus tasks, the first trial was generated using the parameters above, while for trials $i \geq 2$, we used the confidence response on trial $i - 1$ (equivalent to the first task) to adjust the confidence criteria on trial $i$ (equivalent to the second task). We fitted the model to the average confidence autocorrelation value across the two tasks but also examined the obtained confidence autocorrelation within each task.

## Results

### *Experiment 1*

Average accuracy in Experiment 1 was 68% (*SD* = 6.1%) for the letter-identity task. For the color task, average accuracy was 68% (*SD* = 8.6%) in the difficult condition and 88% (*SD* = 10.5%) in the easy condition. Thus, our task was successful in inducing appropriate behavioral performance. Achieving a prespecified level of performance was not critical in this study.

To check for confidence leak, we first performed, for each observer, a simple trial-by-trial correlation of the confidence ratings from the two tasks. The intertask confidence correlation was positive for 25 of the 26 observers (Fig. 2a) and was significantly positive ($p < .05$) for 21 of the 26 observers. Across the whole group, the average Fisher-transformed correlation coefficient was .23 (95% confidence interval, or CI = [.17, .29]), which is significantly

Downloaded from pss.sagepub.com at GEORGIA TECH LIBRARY on February 12, 2016

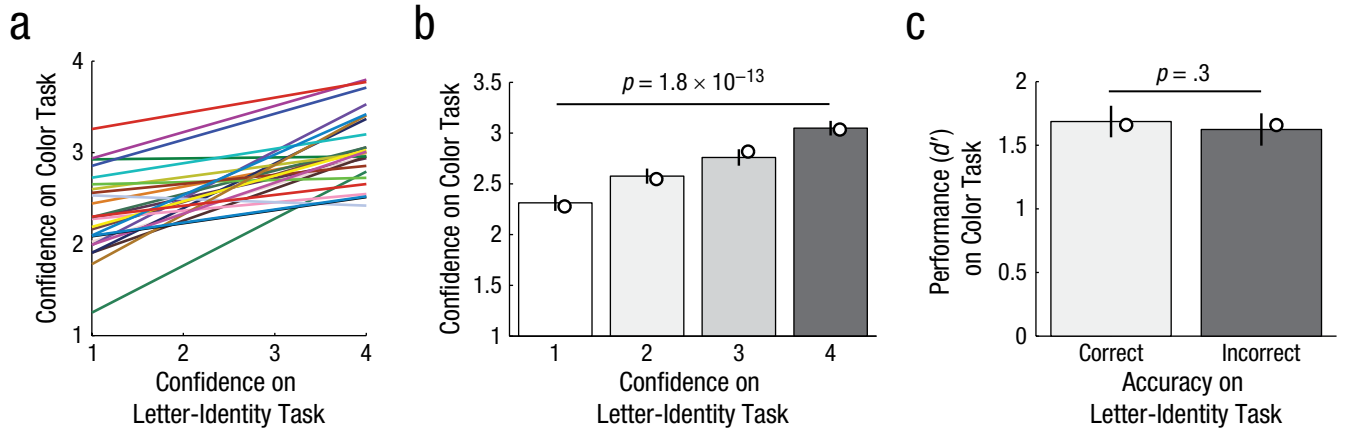


**Fig. 2.** Results of Experiment 1. The best-fitting regression lines in (a) show the relationship between confidence on the letter-identity task and confidence on the color task for each individual observer. Mean confidence on the color task (b) is shown for each level of confidence on the letter-identity task. Mean performance on the color task (c) is shown as a function of accuracy on the letter-identity task. Discs signify model fits, and error bars show standard errors of the mean.

positive, $t(25) = 7.40$, $p = 9 \times 10^{-8}$, and corresponds to an effect size (Cohen's $d = 1.45$) considerably higher than what is conventionally considered large ($d = 0.80$). Finally, the Bayes factor ($9.4 \times 10^9$) revealed substantial evidence for the alternative hypothesis.

Next, to control for the influence of other task factors, we performed a regression in which confidence on the letter-identity task was used to predict confidence on the color task while at the same time controlling for the influence of accuracy and RT on each task, as well as difficulty of the color task (i.e., whether the dominant color was present in 23 or 29 characters). After accounting for these other influences, confidence on the letter-identity task was now a positive predictor of confidence on the color task in all 26 observers. The mean beta value was 0.22 (95% CI = [0.16, 0.27]), $t(25) = 8.02$, $p = 2 \times 10^{-8}$, which signifies that 1 unit of confidence difference in the letter-identity task predicted a 0.22-unit change in confidence in the color task. We also plotted the average confidence on the color task for each confidence rating of the letter-identity task and found a significant difference in their means, $F(3, 25) = 32.12$, $p = 1.8 \times 10^{-13}$ (see Fig. 2b). Notably, neither accuracy nor RT on the letter-identity task predicted confidence on the color task ($p$s = .74 and .22, respectively). Therefore, what leaks is not the quality of the signal itself but observers' metacognitive assessment of the signal correlation.

These data suggest that confidence indeed leaks between tasks, even when these tasks depend on different visual features. However, there are several possible alternative explanations to rule out. First, it could be that observers' attentional states varied over the course of the experiment, which could have led to fluctuations in signal quality that were similar for the two tasks. If signal quality correlates between the two tasks, then confidence would also correlate trivially. To test this possibility, we computed performance on the color task as a function of accuracy on the letter-identity task. After a correct response on the letter-identity task, average $d'$ was 1.69, while after an incorrect response, it was 1.62, with the difference failing to reach statistical significance, $t(25) = 1.03$, $p = .31$, Cohen's $d = 0.10$ (Fig. 2c). Further, this small, nonsignificant performance leak did not significantly correlate across observers with the amount of confidence leak ($r = .19$, $p = .34$). Thus, confidence leak is not simply the result of fluctuations in signal quality because such fluctuations would have produced large corresponding changes in accuracy that would also correlate with confidence leak.

Second, it is possible that confidence ratings for both tasks drifted on a long timescale (e.g., they could have slowly increased for both tasks over the course of the experiment), which would perhaps reflect a change in subjective mood state or degree of perceived expertise with the task. Such low-frequency drift could result in an observed confidence leak even without any within-trial leak. To check for that possibility, we applied scaled correlation (Nikolić, Mureşan, Feng, & Singer, 2012), which is a method to compute the correlation from small window sizes, thus excluding the slow-frequency components from contributing to the correlation. We used window sizes of 5, 10, 20, 50, 100, 200, and 400 trials and found no influence of window size on intertask confidence correlation, $F(6, 25) = 0.64$, $p = .7$, which suggests that confidence leak does not depend on slow changes in the observers' overall cognitive or mood state.

Third, it could be that confidence for both tasks fluctuated on a very short timescale (e.g., less than five trials) and that this fluctuation was similar for both tasks but that there was no causal influence from one task to the

Downloaded from pss.sagepub.com at GEORGIA TECH LIBRARY on February 12, 2016

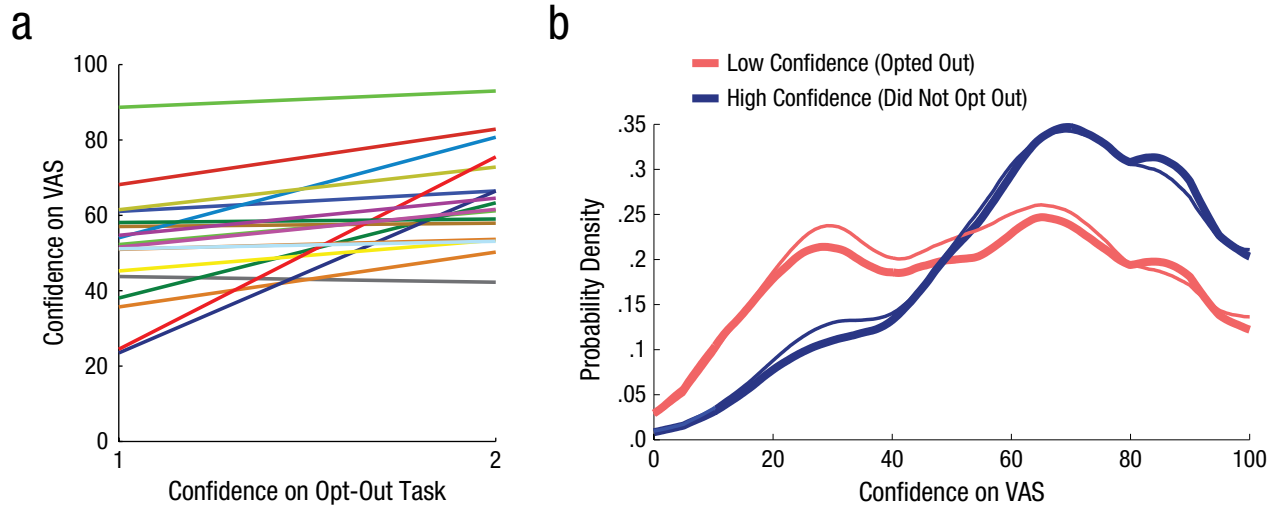


**Fig. 3.** Results of Experiment 2. The best-fitting regression lines in (a) illustrate the relationship between confidence on one task, measured with the visual analogue scale (VAS), as a function of confidence on the other task, measured with the opt-out procedure. Results are shown separately for each individual observer. The probability density functions (b) illustrate confidence ratings on the VAS for one task as a function of observers' confidence on the other task. Confidence was indexed by whether observers opted out of responding (low confidence) or provided a response (high confidence). For display purposes, the curves were smoothed with a 10-point moving average. The thick lines indicate data, and the thin lines indicate model fits.

other. To address this question, we took advantage of a particular feature in our design: While the letter-identity task always had the same difficulty over the course of Experiment 1, the difficulty of the color task changed once every run (of 100 trials). If confidence did not leak from one task to the other (but simply fluctuated simultaneously for the two tasks), we would expect confidence on the letter-identity task to be the same regardless of the difficulty of the color task. However, even though accuracy on the letter-identity task did not depend on the difficulty of the color task ($p$ = .48), observers were significantly more confident on the letter-identity task in the easy color blocks than in the difficult color blocks, $t(25)$ = 2.65, $p$ = .014. That is, an experimental manipulation of task difficulty for the color task led to a significant difference in confidence in the letter-identity task. We obtained a similar result in an additional control experiment in which observers judged the orientation of gray-scale gratings and provided confidence ratings: Intermediate-contrast gratings were judged with higher confidence when presented in the same block as high-contrast gratings than when presented in the same block as low-contrast gratings (Fig. S1 in the Supplemental Material).

Finally, it is possible that observers were subject to motoric priming such that a certain button press used as a confidence response on one task was likely to be repeated for the next task. To exclude simple motoric priming as the cause of the observed results, we analyzed trials in which different buttons were pressed for the two tasks in order to determine whether confidence leak is preserved when trials with the same motoric response are removed. We analyzed the trials in which a confidence rating of 1 or 4 was given on one of the tasks while a different confidence rating was given on the other. Among these trials, a confidence rating of 1 was paired with a confidence of 2, 3, and 4 on 38%, 35%, and 27% of the trials, respectively, while a confidence rating of 4 was paired with a confidence of 1, 2, and 3 on 24%, 30%, and 46% of the trials, respectively. A repeated measures ANOVA confirmed that these patterns are significantly different, $F(2, 19)$ = 4.6, $p$ = .016, which suggests that a simple motoric priming effect cannot explain the data.

## *Experiment 2*

In Experiment 2, we sought to provide stronger evidence that the effects of Experiment 1 were not due to motoric or response priming. We kept the stimulus the same but made the two confidence ratings as different as possible: One of them was provided on a continuous scale, while the other one was collected in an implicit fashion through the opt-out paradigm. We again correlated, for each observer, the confidence ratings from the two tasks on each trial. The intertask correlation was positive for 17 of the 18 observers (Fig. 3a) and was significantly positive ($p$ < .05) for 13 of the 18 observers. Across the whole group, the average Fisher-transformed correlation coefficient was .26 (95% CI = [.10, .42]), which is significantly positive, $t(17)$ = 3.17, $p$ = .006, and corresponds to a large effect size (Cohen's $d$ = 0.75) and strong evidence for the alternative hypothesis (Bayes factor = 13.8). Controlling for accuracy, RT, and difficulty on both tasks

Downloaded from pss.sagepub.com at GEORGIA TECH LIBRARY on February 12, 2016

in a regression analysis (as in Experiment 1) produced virtually identical results. As in Experiment 1, neither accuracy nor RT on one task predicted confidence on the other task (*p*s = .26 and .72, respectively), which suggests that it is specifically the metacognitive assessment of the signal's quality that leaks from one task to another.

To better visualize the magnitude of the effect, we plotted the PDFs of the confidence distributions in the continuous-rating task as a function of confidence on the opt-out task (Fig. 3b). The plot demonstrates the clear tendency for confidence to leak from one task to another: The mean of the VAS PDF for trials in which observers did not opt out was higher than for trials in which they did.

Further, as in Experiment 1, we checked whether the observed confidence leak could be due to correlations in signal quality. We could not compute accuracy for the low-confidence trials in the opt-out task because observers did not provide a guess when they chose to opt out, and we were therefore unable to repeat the analysis from Experiment 1. Instead, we computed the accuracy correlation between the two tasks, restricting the analysis to high-confidence trials on the opt-out task. This correlation ($r$ = .04) was much smaller than the correlation for confidence and did not reach statistical significance ($p$ = .08). Further, as in Experiment 1, this performance leak did not correlate across observers with the amount of confidence leak ($r$ = −.11, $p$ = .67), which confirms that confidence leak is not simply due to fluctuations in signal quality.

To explain the data from Experiments 1 and 2, we created a model of confidence inspired by Bayesian inference. The mathematical description of the model can be found in the Method, and a graphical representation is presented in Figure 4. The crux of the model is that observers attempt to maintain confidence criteria consistently with Bayesian principles, according to which confidence should reflect the likelihood of being correct. However, computing the likelihood of being correct depends on one's prior (which we always fixed to .5 for each choice alternative), the expected likelihood function, and evidence on the current trial. Our model instantiates the intuition that observers interpret a high confidence rating on a previous task as a sign of a high signal-to-noise environment and is thus predictive of likelihoods that are further apart on the current trial. In other words, to account for the expectation of an easier task, confidence criteria are made more liberal. If the expectation is correct, then this adjustment would lead to confidence remaining tightly coupled to a particular accuracy level, as prescribed by normative Bayesian theory. However, if the expectation is incorrect (i.e., high confidence on the previous task does not predict well the difficulty on the current task), this leads to a higher confidence on the current task and therefore to confidence leak. Mathematically, if observers expect the signal-to-noise ratio to change by a factor of $F$, then to remain Bayes optimal in their confidence ratings, they need to shift each of their confidence criteria (defined in the likelihood space) by a factor of $\frac{1}{F}$ (see Equations 10 and 11).

The model employs a single free parameter, θ, that reflects the degree to which observers adjust their expectation of the signal-to-noise ratio on one task as a function of their confidence rating on the other task. Positive values of θ are indicative of positive between-task confidence dependency and therefore indicate the presence of confidence leak. The model provided excellent fit to the data in Experiments 1 and 2.

In Experiment 1, the average value of θ was 0.27, which means that for every increase of confidence by 1 unit (on the scale from 1 to 4) on the letter-identity task, observers expected a signal-to-noise increase of 0.27 on the color task (i.e., a difference of 0.81 between confidence of 1 and 4). Further, θ was positive for 25 of the 26 observers in Experiment 1, $t(25) = 6.42$, $p = 10^{-6}$; Cohen's $d$ = 1.26.

Similarly, θ was positive for 17 of the 18 observers in Experiment 2. The highest value for θ in Experiment 2 was a significant outlier (3.9 *SD* higher than the mean group) and was therefore omitted from the statistical analyses, which demonstrated that θ was significantly positive, $t(16) = 3.83$, $p$ = .001; Cohen's $d$ = 0.93. The average value of θ—with the outlier excluded—was 0.016, which means that for every increase of confidence by 1 unit (on the scale from 0 to 100) on the VAS task, observers expected a signal-to-noise increase of 0.016 on the opt-out task (i.e., a difference of 1.6 between confidence of 0 and 100). Note that the exact value of θ depends on factors such as using the full confidence scale (vs. only a restricted range of the scale). Model fits are shown in Figures 2b and 2c with discs, and in Figure 3b with thin lines. Despite the fact that it had only a single free parameter, the model fit the data in Experiments 1 and 2 very well.

## *Experiment 3*

One critical feature of our model is that confidence leak depends solely on observers' expectation of the signal-to-noise ratio on the current task as a function of the confidence rating on a previous task. This means that the phenomenon of confidence leak should be independent of factors such as the contrast of the stimuli, the individual bias for high or low overall confidence ratings, or the attentional demands of the task. In the Supplemental Material, we report an experiment that demonstrates that confidence leak is indeed independent of stimulus contrast. In Experiment 3, we further tested whether confidence leak depends on attention,

Downloaded from pss.sagepub.com at GEORGIA TECH LIBRARY on February 12, 2016

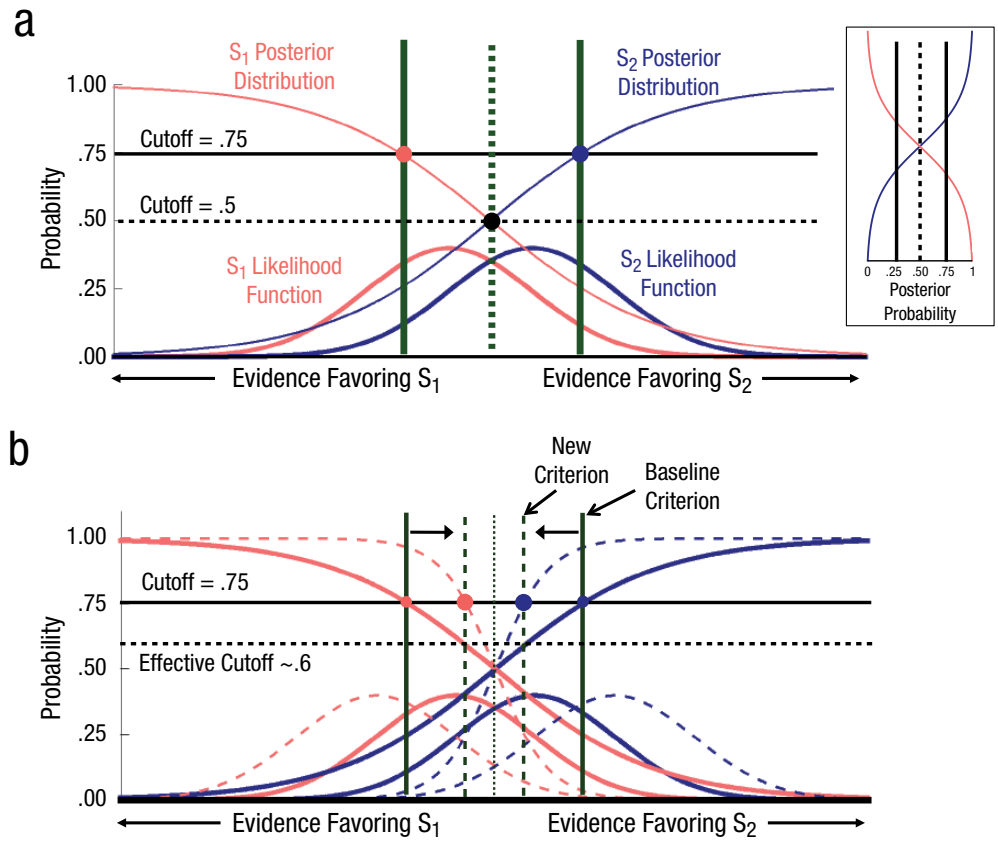


**Fig. 4.** Graphical depiction of our model. Discrimination between stimuli $S_1$ and $S_2$ is shown in (a). Each stimulus follows a Gaussian likelihood function (thick lines) on an axis that denotes the total evidence available on a given trial. The posterior distributions (thin lines) are drawn for cases in which the two stimuli have equal prior probability. Bayes-optimal thresholds (horizontal black lines) are based on predetermined cutoff values (in this case, .5 for the decision and .75 for the confidence). These cutoffs correspond to criteria defined in the likelihood space (vertical lines), which intersect the horizontal thresholds on the posterior distributions. Observers' goal is to set stable cutoffs in the posterior probability space, as depicted in the inset. In (b), solid lines represent the distributions from (a), and dashed lines represent the expected distributions in a higher signal-to-noise environment. For the confidence criteria (defined on the likelihood space) to remain consistent with the threshold placed at a cutoff of .75 on the posterior distributions, they must move "inward" (as illustrated by the difference between the location of the new and baseline criteria). However, if the expectation for a high signal-to-noise environment is false, then the observer is using the criteria indicated by the dashed lines (based on the observer's expectations) to judge stimuli characterized by the distributions indicated by the solid lines. This means that unbeknownst to the observer, he or she is using a lower effective cutoff on the true posterior distribution (in this example, the effective cutoff is ~.6), which naturally results in an increased proportion of high-confidence responses.

which was manipulated by presenting two stimuli (high-attention condition) or four stimuli (low-attention condition; see Fig. 5a).

We first confirmed that higher attention led to better performance, as measured by $d'$. Indeed, $d'$ was higher in the high-attention condition (average $d' = 1.73$) than in the low-attention condition (average $d' = 1.25$), $t(19) = 6.41$, $p = 3.8 \times 10^{-6}$ (Fig. 5b). We then checked for confidence leak by determining the amount of confidence autocorrelation. The assumption in this analysis was that confidence on a given trial would leak to the confidence on the subsequent trial. The confidence autocorrelation was positive for all 20 observers and significant for 18 of them (Fig. 5c). Across the whole group, the average Fisher-transformed correlation coefficient was .28 (95% CI = [.21, .36]), which is significantly positive, $t(19) = 7.66$, $p = 3.2 \times 10^{-7}$, and corresponds to a very large effect size (Cohen's $d = 1.71$). The data also provide strong evidence for the alternative hypothesis (Bayes factor = $2.1 \times 10^{10}$). Controlling for accuracy and difficulty in a regression analysis (as in Experiments 1 and 2) produced virtually identical results.

Downloaded from pss.sagepub.com at GEORGIA TECH LIBRARY on February 12, 2016

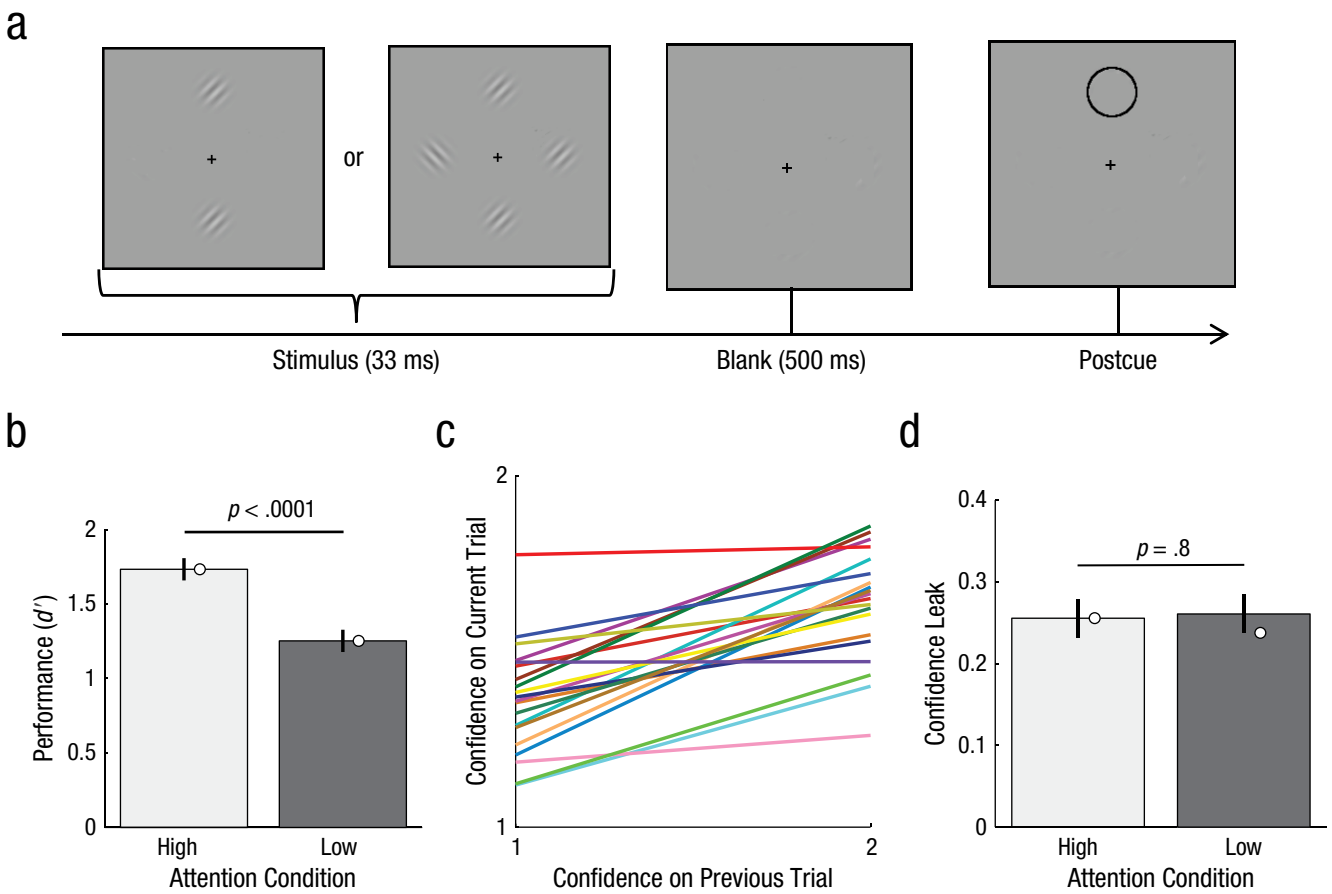


**Fig. 5.** Paradigm and results of Experiment 3. On each trial (a), observers were briefly shown either two (high-attention condition) or four (low-attention condition) Gabor patches, one of which was subsequently postcued. Observers indicated the orientation (clockwise vs. counterclockwise) of the postcued Gabor patch. Performance on the task (b) is shown as a function of attention condition. In (c), best-fitting regression lines show the relationship between average confidence on a given trial and average confidence on the previous trial, separately for each individual observer. Confidence leak (d) is shown as a function of attention condition. In (b) and (d), discs signify model fits, and error bars show standard errors of the mean.

Critically, we compared the amount of confidence leak as defined by the beta value in a regression in which confidence on a given trial was used to predict confidence on the subsequent trial, while controlling for the accuracy and contrast on each trial. We found no significant difference in the beta value in the high-attention (average $\beta = 0.255$) and low-attention (average $\beta = 0.261$) conditions, $t(19) = 0.22$, $p = .83$ (Fig. 5d), which indicates that confidence leak does not depend on attention.

Our model was able to fit these data, too (Figs. 5b and 5d). The model fit indicated that the amount of confidence leak, as defined by the parameter $\theta$, was positive for 19 of the 20 observers, $t(19) = 7.65$, $p = 3.2 \times 10^{-7}$; Cohen's $d = 1.71$, and the average value of $\theta$ was 0.61, which suggests that increasing confidence from low to high led observers to expect a signal-to-noise increase of 0.61 on the next trial.

According to our model, the confidence-leak phenomenon depends on the extent to which a confidence rating on one task or trial shifts observers' expectation for the signal-to-noise ratio on a different task or trial. However, since the environment was kept stable in our experiments (the difficulty of the tasks on any trial could not be predicted from the previous trial), an adjustment in expectation actually led to larger noise in observers' confidence ratings. In our model, larger amounts of confidence leak in individual observers led to larger erroneous shifts in the likelihood functions, which, in turn, should result in lower metacognition scores (Maniscalco & Lau, 2012). We checked for such an effect by computing, for each observer, the Type 2 AUC (Fleming & Lau, 2014; Fleming et al., 2010). We then correlated this measure of metacognition with the amount of confidence leak, as estimated by the Fisher-transformed intertask correlation value. To increase power, we performed the correlation

Downloaded from pss.sagepub.com at GEORGIA TECH LIBRARY on February 12, 2016

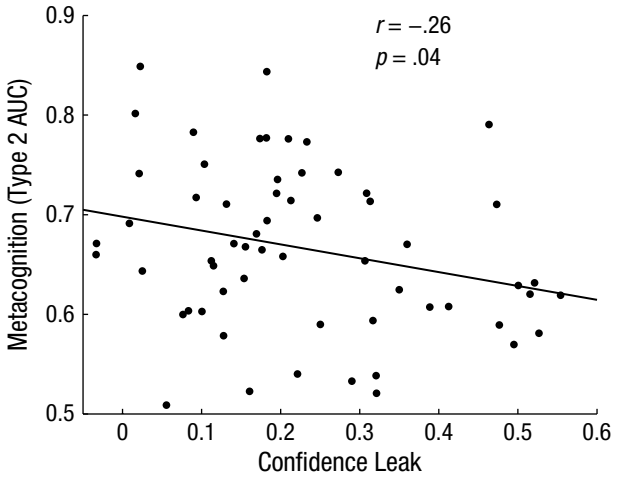


**Fig. 6.** Scatterplot (with best-fitting regression line) showing the relationship between mean metacognition score and confidence leak. Metacognition was measured as the area under the Type 2 receiver-operating-characteristic curve (Type 2 AUC).

on the combined data from Experiments 1 through 3. We found that confidence leak correlated negatively with metacognition ($r = -.36$, $p = .004$, 95% CI = [−.56, −.12]). The correlation appeared to be influenced by 4 outliers (2 observers with Fisher-transformed correlations > 1 and 2 observers with metacognitive scores worse than chance, where chance = .5); however, this finding remained significant when we instead applied Spearman's rank correlation, which is less sensitive to outliers ($\rho = -.28$, $p = .02$). We nonetheless confirmed that the correlation remained significantly negative ($r = -.26$, $p = .04$) when we removed the outliers altogether (Fig. 6). Notably, when the same analysis was performed within each experiment, we obtained similar correlation magnitudes ($r$s = −.31, −.45, and −.26, respectively) but because of the smaller number of observers in each analysis, the $p$ values did not reach the .05 level ($p$s = .12, .06, and .27, respectively). Taken together, these results show that, as suggested by our model, the phenomenon of confidence leak impairs observers' metacognitive ability to introspect on their accuracy.

### *Experiment 4*

The empirical relationship between confidence leak and metacognitive accuracy led us to explore whether the two may share a common neurophysiological representation. We reanalyzed the data from McCurdy et al. (2013), who replicated a finding first reported by Fleming et al. (2010) that higher metacognitive sensitivity is related to higher gray matter volume in the aPFC. Analyzing the data from Experiment 4, we found, behaviorally, very strong evidence for confidence leak in that data set, too: Confidence autocorrelation was positive for all 34 observers, $M = .27$, 95% CI = [.22, .31], $t(33) = 10.9$, $p = 1.9 \times 10^{-12}$, Cohen's $d = 1.87$, Bayes factor = $2.4 \times 10^{24}$. In addition, replicating the analysis on the correlation between confidence leak and metacognitive scores in Experiments 1 through 3, we found that confidence leak correlated negatively ($r = -.38$, $p = .025$) with the measure of metacognitive efficiency (meta $d'/d'$; Maniscalco & Lau, 2012) originally used in the analyses reported by McCurdy et al. (2013).

We then explored whether confidence leak is related to gray matter volume. Given the negative relation between confidence leak and metacognitive scores, we expected a negative relationship between confidence leak and gray matter volume in PFC regions that had previously been linked to metacognitive sensitivity (Fleming et al., 2010). We used the same methods as McCurdy et al. (2013) and found two regions in right PFC for which lower gray matter volume predicted higher confidence-leak scores. Both regions survived small-volume correction for multiple comparisons—peak voxel coordinates for right dorsolateral PFC: $x = 41$, $y = 32$, $z = 22$; $t = 3.76$, cluster-level familywise-error-corrected (FWE-corrected) $p = .032$; peak voxel coordinate for right aPFC: $x = 35$, $y = 53$, $z = 6$; $t = 3.71$, cluster-level FWE-corrected $p = .031$ (Fig. 7).

As we explained previously, confidence leak logically leads to decreased metacognitive scores. However, it is also possible that observers with lower metacognitive capacity are more susceptible to confidence leak (i.e., the reverse causal link). Such an effect would raise the possibility that the PFC regions identified here do not contribute directly to confidence leak. Therefore, we regressed out the influence of metacognitive scores from the confidence-leak scores and repeated the analyses with the residuals. We found that these new confidence-leak scores were still significantly predicted by lower gray matter volume in the region in right dorsolateral PFC ($t = 4.46$, cluster-level FWE-corrected $p = .026$) but not by the region in aPFC ($p > .05$).

## Discussion

Four experiments provided evidence for the novel phenomenon of confidence leak between different psychophysical tasks. In the first two experiments, confidence leaked between the tasks of judging the dominant letter identity and color of two types of stimuli. In the last two experiments, confidence leaked over time in the same task. Our work extends previous demonstrations of serial dependence in perceptual decisions (Fischer & Whitney, 2014; Fründ et al., 2014; Liberman et al., 2014; Zhang et al., 2014) and reveals that such dependence is also present for metacognitive judgments.

Downloaded from pss.sagepub.com at GEORGIA TECH LIBRARY on February 12, 2016

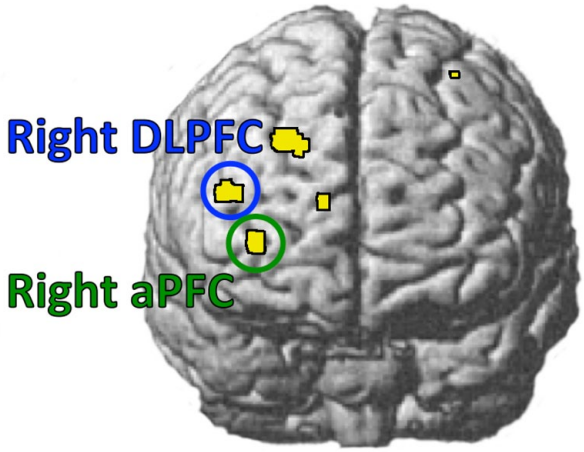


**Fig. 7.** Brain correlates of confidence leak: regions in which higher confidence-leak scores were predicted by lower gray matter volume. The image shows a *t* map with a threshold of $p = .005$, uncorrected, for display purposes, though analyses were performed using small-volume correction for multiple comparisons. Regions with significant activations are circled. aPFC = anterior prefrontal cortex; DLPFC = dorsolateral prefrontal cortex.

What causes serial dependence in perception? Previous work has theorized the existence of a continuity field that arises from the brain's attempt to use previous information to interpret the current visual scene (Fischer & Whitney, 2014; Liberman et al., 2014; Yu & Cohen, 2009). This body of work has shown that in the case of percepts, this effect is likely to be perceptual rather than cognitive. The metacognitive continuity field that we have demonstrated appears to arise from a similar process: The brain attempts to interpret the fidelity of its current decision based on the perceived fidelity of past decisions. On the other hand, the phenomenon of confidence leak clearly arises from higher-level, rather than perceptual, processes, as demonstrated by the fact that it appears across perceptual tasks and is dependent on the brain anatomy of the prefrontal cortex.

Our computational model formalizes the idea that confidence on a previous task is used to predict the quality of the perceptual signal in a current task. Using the principles of Bayesian inference, the model describes normative behavior in cases in which observers expect that the environment has a predictable higher-order structure (David et al., 2004) that can be exploited (Fischer & Whitney, 2014; Fründ et al., 2014; Liberman et al., 2014). Specifically, it has been argued that observers assume the world is autocorrelated because it usually is and that this assumption is not particularly detrimental to performance when the world is not autocorrelated (Yu & Cohen, 2009). The model specifies precisely how expecting higher or lower signal strength on a given trial ought to translate into an adjustment of the metacognitive criterion for confidence. Despite employing a single free parameter, the model fit a large amount of data well (see Figs. 2, 3, and 5). It also made two novel predictions that were confirmed by additional experiments or analyses. First, the model predicted that the phenomenon of confidence leak does not depend on factors that affect the perceptual signal, such as attention and contrast, which was confirmed in Experiment 3 and the control experiment reported in the Supplemental Material, respectively. Second, the model predicted that stronger confidence leak leads to lower metacognitive sensitivity (Fleming & Lau, 2014), which was confirmed with the combined data from Experiments 1 through 3, as well as in Experiment 4. Thus our model not only fit the existing data well but also made testable novel predictions.

Several components of our model may seem counterintuitive. For example, why do observers use one task to predict their performance on a completely different task? We think that this is because in the real world, even very different tasks tend to vary in difficulty together. For instance, foggy conditions, headache, or a high level of distractibility could decrease the signal-to-noise ratio similarly for both the color and letter-identity tasks. Another potential concern is whether the fact that highly confident responses make people more likely to give higher confidence on the next trial or task would lead to confidence increasing indefinitely. We note that confidence ratings are inherently based on both signal strength and perceptual expectations (as well as other potential factors). Thus, even after a maximal confidence rating, observers can still have a relatively low confidence rating on the next task or trial if the signal strength happens to be relatively low. Finally, if observers are constantly making predictions about the likelihood functions of the upcoming tasks or trials, why do they not learn to adjust these predictions over time? Because the expected likelihood functions were only modestly shifted compared with the real likelihood functions, most trials would likely not provide enough evidence to overturn observers' expectation for an autocorrelated environment, especially in the absence of trial-to-trial feedback. In a similar manner, purely perceptual decisions remain autocorrelated despite the perceptual conflict this autocorrelation brings (Fischer & Whitney, 2014; Fründ et al., 2014; Liberman et al., 2014; Zhang et al., 2014).

Our results have several important implications. First, the fact that confidence leak predicts metacognitive sensitivity means that we have identified one of the causes of suboptimal metacognitive performance in perceptual decision making (Fleming et al., 2010; Maniscalco & Lau, 2012; McCurdy et al., 2013). Second, our findings provide evidence that confidence for perceptual decisions is not solely determined by the signal strength in visual cortex,

Downloaded from pss.sagepub.com at GEORGIA TECH LIBRARY on February 12, 2016

as assumed in some dominant theories. For example, theories of population coding in early visual cortex (Ma, Beck, Latham, & Pouget, 2006) postulate that visual cortex activity forms a distribution from which one can directly read the decision (usually the peak of the distribution) and the uncertainty (the width of the distribution; Drugowitsch & Pouget, 2012). However, if confidence were determined exclusively on the basis of activity in the visual cortex (and perhaps additionally corrupted by white noise), then we should not have seen evidence for confidence leak in Experiments 1 and 2. Indeed, the two tasks in these experiments focused on different visual features (color and letter identity), which are often assumed to rely on processing in largely independent neural populations (James, James, Jobard, Wong, & Gauthier, 2005; Zeki, 1990). Notably, even for neurons that may be sensitive to both shape and color, it is hard to imagine how and why sharper representation for one of these features (corresponding to high confidence) could lead to a sharper representation for the other feature.

The phenomenon of confidence leak extends previous work by De Gardelle and Mamassian (2014), who recently demonstrated the existence of a "common currency" for confidence in which the certainty for different perceptual tasks can be directly and meaningfully compared. This finding suggests that confidence is represented in generic, task-independent format, which is a necessary condition for our proposed mechanism for confidence leak. Our results also build on previous work by Mueller and Weidemann (2008), who found confidence autocorrelation in the same task but did not put forward an explanation of the causes of such confidence autocorrelation or show dependence between different tasks. Finally, our results are consistent with those of previous studies suggesting that confidence is influenced by a subjective perception of ease (Koriat, 2008, 2011), but they extend this previous literature by considering between-task influences.

One alternative explanation for our results is that providing confidence on one task simply changes the emotional state of the observer. This explanation is not mutually exclusive with the theory that observers engaged in Bayesian inference: For example, it is possible that an expectation of higher signal quality puts observers in a better mood. Nevertheless, we consider a purely emotional account of the confidence-leak phenomenon to be unlikely. Indeed, in Experiment 2, we gave observers the opportunity to earn points and presented them with the opt-out paradigm in which they could choose not to respond if they were unsure about their response. Further, we explicitly informed them that the optimal strategy was to choose the opt-out option when they were less than 66% certain about their response. Observers were promised a monetary reward for high performance, and it is therefore likely that they tried to minimize purely emotional reactions and instead gave their response based on the objectively computed probability of being correct. Despite that, in Experiment 2, we observed confidence leak that was just as strong as in the other experiments, which suggests that a purely emotional account is unlikely.

An important question for future research relates to the limit of the confidence-leak phenomenon. Future studies should examine whether confidence leak occurs in tasks that depend on different senses altogether, such as vision and audition. Beyond perceptual tasks, an important question is whether confidence ratings on any task, such as memory or high-level cognitive tasks, also show confidence leak. If similar confidence leak is present for cognitive tasks, there will likely be a number of real-world implications related, for example, to how the level of confidence of witnesses in courts, doctors examining mammograms, or drivers deciding whether road conditions are dangerous can be influenced by seemingly irrelevant contexts.

## Author Contributions

D. Rahnev and H. Lau developed the study concept and design. Testing and data collection were performed by D. Rahnev, A. Koizumi, and L. Y. McCurdy. D. Rahnev analyzed and interpreted the data under the supervision of H. Lau and M. D'Esposito. D. Rahnev drafted the manuscript, and H. Lau, A. Koizumi, L. Y. McCurdy, and M. D'Esposito provided critical revisions. All authors approved the final version of the manuscript for submission.

## Acknowledgments

We thank Sneha Subramanian for helpful comments and Tashina Graves for collecting the data for Experiment 3.

## Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

## Funding

This work was supported by the National Institute of Health (Grant No. R01 NS088628-01) and the John Templeton Foundation (Grant No. 21569).

## Supplemental Material

Additional supporting information can be found at http://pss.sagepub.com/content/by/supplemental-data

## Open Practices

All data and materials have been made publicly available via Zenodo and can be accessed at http://dx.doi.org/10.5281/zenodo.18545. The complete Open Practices Disclosure for

Downloaded from pss.sagepub.com at GEORGIA TECH LIBRARY on February 12, 2016

this article can be found at http://pss.sagepub.com/content/by/supplemental-data. This article has received badges for Open Data and Open Materials. More information about the Open Practices badges can be found at https://osf.io/tvyxz/wiki/1.%20View%20the%20Badges/ and http://pss.sagepub.com/content/25/1/3.full.

## References


Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.

David, S. V., Vinje, W. E., & Gallant, J. L. (2004). Natural stimulus statistics alter the receptive field structure of v1 neurons. *The Journal of Neuroscience*, *24*, 6991–7006. doi:10.1523/JNEUROSCI.1422-04.2004

De Gardelle, V., & Mamassian, P. (2014). Does confidence use a common currency across two visual tasks? *Psychological Science*, *25*, 1286–1288. doi:10.1177/0956797614528956

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. London, England: Palgrave Macmillan.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, Article 781. Retrieved from http://journal.frontiersin.org/article/10.3389/fpsyg.2014.00781/full

Drugowitsch, J., & Pouget, A. (2012). Probabilistic vs. non-probabilistic approaches to the neurobiology of perceptual decision-making. *Current Opinion in Neurobiology*, *22*, 963–969. doi:10.1016/j.conb.2012.07.007

Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nature Neuroscience*, *17*, 738–743. doi:10.1038/nn.3689

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: Computation, biology and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 1280–1286. doi:10.1098/rstb.2012.0021

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, Article 443. Retrieved from http://journal.frontiersin.org/article/10.3389/fnhum.2014.00443/full

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. doi:10.1126/science.1191883

Fründ, I., Wichmann, F. A., & Macke, J. H. (2014). Quantifying the effect of intertrial dependence on perceptual decisions. *Journal of Vision*, *14*(7), Article 9. Retrieved from http://jov.arvojournals.org/article.aspx?articleid=2194025

James, K. H., James, T. W., Jobard, G., Wong, A. C. N., & Gauthier, I. (2005). Letter processing in the visual system: Different activation patterns for single letters and strings. *Cognitive, Affective, & Behavioral Neuroscience*, *5*, 452–466.

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

Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Davies (Eds.), *Cambridge handbook of consciousness* (pp. 289–326). New York, NY: Cambridge University Press.

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 945–959. doi:10.1037/0278-7393.34.4.945

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

Liberman, A., Fischer, J., & Whitney, D. (2014). Serial dependence in the perception of faces. *Current Biology*, *24*, 2569–2574. doi:10.1016/j.cub.2014.09.025

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

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

Maniscalco, B., Bang, J. W., Iravani, L., Camps-Febrer, F., & Lau, H. (2012). Does response interference depend on the subjective visibility of flanker distractors? *Attention, Perception, & Psychophysics*, *74*, 841–851. doi:10.3758/s13414-012-0291-2

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

McCurdy, L. Y., Maniscalco, B., Metcalfe, J., Liu, K. Y., de Lange, F. P., & Lau, H. (2013). Anatomical coupling between distinct metacognitive systems for memory and visual perception. *The Journal of Neuroscience*, *33*, 1897–1906. doi:10.1523/JNEUROSCI.1890-12.2013

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. doi:10.3758/PBR.15.3.465

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and some new findings. In G. Bower (Ed.), *The psychology of learning and motivation* (pp. 125–141). New York, NY: Academic Press.

Nikolić, D., Mureşan, R. C., Feng, W., & Singer, W. (2012). Scaled correlation analysis: A better way to compute a cross-correlogram. *The European Journal of Neuroscience*, *35*, 742–762. doi:10.1111/j.1460-9568.2011.07987.x

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

Rahnev, D., 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. doi:10.1152/jn.00184.2012

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. doi:10.1152/jn.00209.2013

Rahnev, D., Lau, H., & De Lange, F. P. (2011). Prior expectation modulates the interaction between sensory and prefrontal regions in the human brain. *The Journal of Neuroscience*, *31*, 10741–10748.


Downloaded from pss.sagepub.com at GEORGIA TECH LIBRARY on February 12, 2016

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. doi:10.1038/nn.2948

Rahnev, D., 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. doi:10.1152/jn.00985.2011

Shimamura, A. P. (2000). Toward a cognitive neuroscience of metacognition. *Consciousness and Cognition*, *9*, 313–326. doi:10.1006/ccog.2000.0450

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

Yu, A. J., & Cohen, J. D. (2009). Sequential effects: Superstition or rational behavior? *Advances in Neural Information Processing Systems*, *21*, 1873–1880.

Zeki, S. (1990). A century of cerebral achromatopsia. *Brain*, *113*, 1721–1777.

Zhang, M., Wang, X., & Goldberg, M. E. (2014). A spatially nonselective baseline signal in parietal cortex reflects the probability of a monkey's success on the current trial. *Proceedings of the National Academy of Sciences, USA*, *111*, 8967–8972. doi:10.1073/pnas.1407540111

Downloaded from pss.sagepub.com at GEORGIA TECH LIBRARY on February 12, 2016