Running Head: Expectations improve perceptual metacognition

Prior expectations facilitate metacognition for perceptual decision

Sherman, M.T.[1,2 *]

Seth, A.K.[2,3]

Barrett, A.B. [2,3]

Kanai, R.[1,2]

[1] School of Psychology, University of Sussex, Falmer, BN1 9QH, UK

[2] Sackler Centre for Consciousness Science, University of Sussex, Falmer, BN1 9QJ, UK

[3]Department of Informatics, University of Sussex, Falmer, BN1 9QJ, UK

***To whom correspondence should be addressed:** Maxine T. Sherman, Email: M.Sherman@Sussex.ac.uk, Permanent address: School of Psychology, Pevensey Building, University of Sussex, Falmer, BN1 9QH, UK

Abstract

The influential framework of 'predictive processing' suggests that prior probabilistic expectations influence, or even constitute, perceptual contents. This notion is evidenced by the facilitation of low-level perceptual processing by expectations. However, whether expectations can facilitate high-level components of perception remains unclear. We addressed this question by considering the influence of expectations on perceptual metacognition. To isolate the effects of expectation from those of attention we used a novel factorial design: expectation was manipulated by changing the probability that a Gabor target would be presented; attention was manipulated by instructing participants to perform or ignore a concurrent visual search task. We found that, independently of attention, metacognition improved when yes/no responses were congruent with expectations of target presence/absence. Results were modeled under a novel Bayesian signal detection theoretic framework which integrates bottom-up signal propagation with top-down influences, to provide a unified description of the mechanisms underlying perceptual decision and metacognition.

Keywords: Metacognition; Expectation; Decision Making; Perceptual Confidence; Attention

1. Introduction

Metacognition, or 'cognition about cognition', reflects the knowledge we have of our own decision accuracy and comprises an important, high-level component of decision making in both perceptual and cognitive settings. In perceptual decision, metacognition is often operationalized as the trial-by-trial correspondence between (objective) decision accuracy and (subjective) confidence. A key question in perceptual metacognition is how, and indeed whether, metacognition is affected by top-down influences, such as attention and expectation. In the case of attention, it has long been known that it can improve visual target detection (Posner, 1980). However, the relationship between attention, confidence, and metacognition remains unclear. While Kanai and colleagues found that perceptual metacognition persists when attention is diverted (Kanai, Walsh, & Tseng, 2010), other studies suggest that the absence of attention can lead to overconfidence (Rahnev et al., 2011; Wilimzig & Fahle, 2008).

Inspired by the growing influence of 'predictive processing' or 'Bayesian brain' approaches to perception and cognition (for reviews, see Clark, 2013; Summerfield & de Lange, 2014), empirical work on top-down attention is now complemented by a growing focus on the role of top-down expectations in decision making. In Bayesian terms, expectations can be conceived as prior beliefs that constrain the interpretation of sensory evidence. It has been shown that prior knowledge, either of stimulus timing ('when') or of stimulus features ('what'), facilitates low-level processing, as reflected in measures such as reaction time (Stefanics et al., 2010) and contrast sensitivity (Wyart, Nobre, & Summerfield, 2012). Such improvements are often accompanied by the attenuation of both the BOLD responses and ERP amplitude

following expected relative to unexpected perceptual events (Egner, Monti, & Summerfield, 2010; Melloni, Schwiedrzik, Müller, Rodriguez, & Singer, 2011; Wacongne et al., 2011). As well as facilitating low-level perception, expectations may influence conscious content. This idea is supported by evidence for expectations inducing subjective directionality in ambiguous motion (Sterzer, Frith, & Petrovic, 2008) and lowering the threshold of subjective visibility for previously seen versus novel visual stimuli (Melloni et al., 2011). These effects are similar to those exerted by top-down attention. However, while it has been argued that attention and expectation reflect similar processes (Desimone & Duncan, 1995; Duncan, 2006), orthogonal manipulations of attention and expectation have demonstrated that, although they are tightly intertwined, they can have separable effects on neural activity (Hsu, Hämäläinen, & Waszak, 2014; Jiang, Summerfield, & Egner, 2013; Kok, Rahnev, Jehee, Lau, & de Lange, 2012; Wyart et al., 2012).

One influential process theory within the Bayesian brain framework is predictive coding (Beck et al., 2009; Clark, 2013; Friston, 2009; J. Hohwy, 2013; Lee & Mumford, 2003). Predictive coding also posits that efficient processing is achieved by constraining perceptual inference according to the prior likelihood of that inference ('expectations'). Here, the predictive models underlying perception are generally assumed to be multilevel and hierarchical in nature (Clark, 2013; Friston, Adams, Perrinet, & Breakspear, 2012), incorporating priors related both to low-level stimulus features, and to high-level features representing object-level invariances. Plausibly, priors concerning subjective confidence for perceptual decisions may be implemented at high levels of the hierarchy. Based on this possibility, we set out to investigate whether the top-down influences of attention and prior expectation modulate perceptual metacognition.

To address whether expectation can improve metacognition we orthogonally manipulated both attention and expectation. This separated their effects, and was achieved by adopting a dual-task design. In a Gabor detection task, expectation was manipulated by informing participants of the probability of Gabor presence or absence as it changed over blocks. In this way, certain blocks induced an expectation of Gabor presence and others, of absence. In half of the blocks, participants were instructed to additionally perform a concurrent visual search task that diverted attention away from the detection task.

Objective performance can be assessed by using type 1 signal detection theory (SDT). By comparing signal type (e.g. present, absent) and response (present, absent), type 1 SDT enables a computation of independent measures of objective sensitivity and decision threshold (*d'* and *c*, respectively). We used type 2 SDT to assess metacognitive sensitivity. By obtaining trial-by-trial retrospective confidence ratings, metacognitive sensitivity and confidence thresholds can be computed from response accuracy and decision confidence. We used two such methods – type 2 *D'*, which is a direct analogue of type 1 *d'* (Kunimoto, Miller, & Pashler, 2001), and meta-*d'* (see Section 2.5.2 or Barrett, Dienes, & Seth, 2013; Maniscalco & Lau, 2012; Rounis, Maniscalco, Rothwell, Passingham, & Lau, 2010). Given that prior expectations have been shown to facilitate low-level processing, we hypothesized that expectations would also improve metacognitive sensitivity. We tested this hypothesis by considering the congruency between participants' yes/no decision and the block-wise expectation of Gabor presence or absence. Specifically, we hypothesized that metacognitive sensitivity would be greater following expectation-congruent type 1 decisions (e.g. reporting target presence when

expecting target presence), than following expectation-incongruent decisions (e.g. reporting presence when expecting absence).

2. Methods

2.1 Participants

Twenty-one participants (14 female) completed the experiment. All were healthy students from the University of Sussex, aged 18 to 31 (*M* = 20.4, *SD* = 3.2) and had normal or corrected-to-normal vision. The sample size for adequate power was computed using GPower 3.1 (Faul, Erdfelder, Lang, & Buchner, 2007), with estimated effect sizes derived from pilot studies. Data from one participant were excluded because their visual search task performance deviated by more than 1.5 *SD* from the mean (98.6% correct) and another, for having no variability in their confidence reports (100% confident). This left data from 19 participants for analysis, all of whom demonstrated, averaging over conditions, a Gabor detection *d'* and visual search accuracy that was within 1.5 *SD* from the mean. Participants were offered course credits for participating and informed, written consent was obtained. The experiment was approved by the University of Sussex ethics committee.

2.2 Stimuli and setup

Stimuli were generated using the Psychophysics toolbox for Matlab (Brainard, 1997; Kleiner, D., & Pelli, 2007) and presented on a 20 inch Dell Trinitron CRT display (resolution 1048x768; refresh rate 85 Hz). Participants were tested individually in darkened rooms and were seated 60cm away from the screen. Both stimuli and background were linearised using a Minolta LS-100 photometer (γ = 2.23607, Weibull fit). The background was grayscale and uniform.

2.3. Design and procedure.

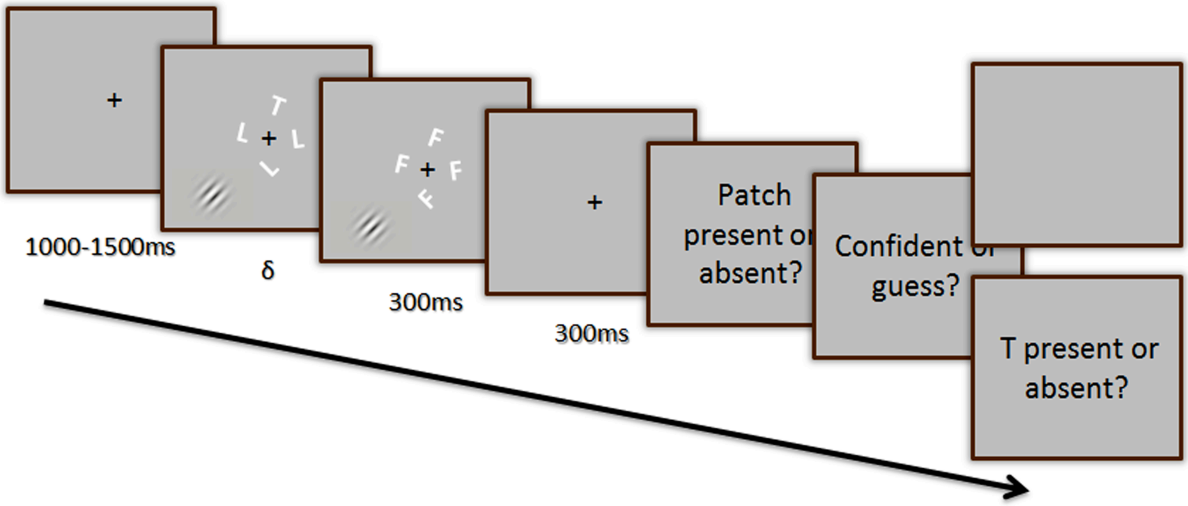

Figure 1. Trial sequence across both staircases and experimental trials. In this trial, both the visual search and detection targets (T and Gabor, respectively) are present. Participants are prompted to respond to the visual search display in diverted attention trials (final, bottom) but not full attention trials (final, top). δ signifies the time that the visual search Ls and Ts were presented for. This time was titrated for each participant individually.

This experiment implemented a novel dual-task design, which is depicted in figure 1. The critical task was to report the presence or absence of a near-threshold Gabor patch (which indeed, was either present or absent). The second task was a visual search task, in which it had to be determined whether a target (the letter 'T'), had been present or absent amongst distracters (letter 'L's).

Trials began with the presentation of a white central fixation cross (0.38°x 0.38°, random duration between 500 and 1,500 ms). This was followed, on Gabor

present trials only, by the appearance of the peripheral Gabor patch (spatial frequency 2c/°, Gaussian *SD* = 2°) in the lower-right quadrant of the screen. On each presentation, the phase was either 45° or 225° (50% chance of each). To reduce sensory adaptation effects, the precise location in which it was presented was jittered in both the horizontal and vertical direction from a baseline position of 25.2° x 21.08°. On each trial the jitter for each direction was randomly sampled from the interval [0.66°, 1.24°]. The contrast of the Gabor was titrated for each participant so that hit rate was 79.4% (see section 2.4, Staircases). In total it was presented for 388ms and had a gradual onset and offset.

Immediately following the offset of the fixation cross, the central visual search array also appeared. On Gabor present trials, the Gabor and the array were therefore presented simultaneously. The array consisted of four white letters (1.43° x 1.43°) – either 3 'L's and a 'T' (visual search target present, 50% chance) or 4 'L's (visual search target absent, 50% chance) - arranged around fixation at 0°, 90°, 180° and 270°. Trial-by-trial, the orientation of each letter took a random value between 0° and 359°. The time for which the letters remained on-screen was adjusted for each participant so that visual search percent correct was 79.4% (*M* = 254 ms, *SD* = 75 ms. See section 2.4, Staircases). To ensure that the task was difficult enough to divert attention, the array of letters was backwards-masked by an array of 'F's that remained on screen for 300ms. This masking array was followed by a series of on-screen response prompts, requesting un-speeded, key-press responses to: first, the Gabor task (Gabor present or absent); second, binary confidence in the accuracy of that report (confident or guess); finally, and in diverted attention conditions only (see next paragraph), the visual search task (T present or absent).

Expectation was manipulated in the Gabor task by changing the probability that it would be present versus absent over blocks of trials (25%, 50% or 75% probability of target presence). In the 25% condition, where Gabor presence was unlikely, an expectation of absence was induced. The 50% condition was a control, and in the 75% condition an expectation of presence was induced. Orthogonally to this expectation manipulation, attention was manipulated over blocks of trials by instructing participants to either perform or ignore the concurrent visual search task. When participants were in a 'perform visual search' block, their attention was diverted from the critical Gabor detection task, whereas when they were instructed to ignore the visual search array, their attention was fully focused on Gabor detection. In the diverted attention condition, participants were instructed to prioritize the visual search task. Thus, each block was associated with an expectation of Gabor presence or absence and a degree of attentional resource for the Gabor task (full/diverted). Before each block began, both the probability of Gabor presentation and instructions to either perform both tasks or ignore the Gabor were presented on-screen. At the end of each block, if visual search accuracy had dropped below 60% on-screen feedback reminded participants to maintain their concentration on the visual search task. Participants completed 36 blocks in total (6 of each of the 6 conditions, counterbalanced) and each block had 12 trials. This gave a total of 432 trials.

Before data collection began, instructions for the tasks were presented on-screen. The on-screen instructions were additionally read to the participant to ensure that they were fully understood. These explained that the probability of target presentation in the upcoming block would be given (25%, 50% or 75%) and that the information was correct and would help them complete the difficult task. Participants

were instructed to fixate centrally throughout and to be as accurate as possible in all of their (un-speeded) responses. Next, participants completed a set of practice trials for each type of task (staircases and experimental conditions). Next, three psychophysical staircase procedures were completed (see section 2.4) and finally, the experimental trials. Once all experimental trials had been completed, participants were debriefed.

2.4 Staircases.

We required performance in the Gabor detection task to be equated across levels of attention and across participants. Furthermore, the difficulty of the visual search task also had to be controlled across participants. Before the experimental trials began, three adaptive staircase procedures were therefore completed. The first staircase adjusted Gabor contrast under full attention, the second, the time for which visual search Ls and Ts were presented and the third, Gabor contrast under diverted attention. The staircases set performance (percent correct for the visual search task and hit rate for the Gabor task) in each task at 79.4%. Each of the three procedures consisted of two interleaved, identical staircases which terminated after 8 reversals. The visual display was identical to that in experimental trials (see section 2.3 and figure 1), however the reports requested from participants varied across procedures. During these procedures, confidence judgments were not requested and there was a 50% chance of Gabor presentation.

In staircase 1, Gabor detection was performed under full attention (i.e. ignore visual search). Participants were instructed to fixate centrally, ignore the visual search display and report peripheral Gabor presence or absence. The initial contrast of the attended target Gabor was 5% and this was titrated by the staircases. The

(ignored) visual search Ls and Ts were presented for 300 ms before they were masked.

In staircase 2, the visual search task was performed but the Gabor task was not. Participants were instructed to ignore the Gabor and only perform the visual search task. Here, they reported whether a target T was present or absent in an array of distracter Ls. The visual search array of Ls and Ts were initially presented for 300 ms before being masked, and this duration was titrated by the staircases. The (ignored) Gabor, if present, had the contrast determined in staircase 1.

In staircase 3, both tasks were performed. Participants were instructed to prioritize the visual search task while concurrently performing the Gabor detection task. Visual search letters were presented for the duration determined in staircase 2. The Gabor was initially presented at 1.05 times the contrast level acquired in staircase 1. The contrast of the unattended Gabor was titrated over the course of the procedure. Participants responded to the Gabor task first and the visual search task second (as in the experimental trials). If participants' mean visual search accuracy across the staircase dropped below 60% they received on-screen instructions to maintain concentration on the visual search task.

2.5 Analysis

2.5.1. Statistical analyses. Objective detection performance for the Gabor detection task was assessed using type 1 signal detection theory (SDT; Green & Swets, 1966) measures *d'* (detection sensitivity) and *c* (decision threshold). A negative/positive *c* reflects a bias towards reporting target presence/absence. Visual

search performance was also assessed using *d'* and *c*. Because we required *d'* and *c* values to remain independent of each other, adjusted type 1 *d'* was not used.

Unless otherwise stated, alpha is set at 5% , the assumption of sphericity has been met and post-hoc tests are FDR corrected (Banjamini & Hochberg, 1995) throughout.

2.5.2 Type 2 Signal Detection Theory. Metacognitive sensitivity was measured by obtaining trial-by-trial confidence ratings and using type 2 SDT to assess the relationship between confidence and accuracy (Barrett, Dienes, & Seth, 2013; Galvin, Podd, Drga, & Whitmore, 2003; Kunimoto et al., 2001; Macmillan & Creelman, 2004). Type 2 measures are calculated analogously to the type 1 case: type 2 hits (correct and confident) and correct rejections (incorrect and guess) are compared with type 2 misses (correct and guess) and false alarms (incorrect and confident). From these, type 2 *D'* (metacognitive sensitivity) and type 2 *C* (confidence threshold) can be computed (Kunimoto et al., 2001). Type 2 hit rate (HR) and type 2 false alarm rate (FAR) are calculated as follows (where the subscript '2' indicates type 2 SDT outcomes):

$$HR = \frac{\sum H_2}{\sum H_2 + \sum M_2}, \qquad FAR = \frac{\sum FA_2}{\sum FA_2 + \sum CR_2}$$

Thus, HR reflects confidence for correct responses and FAR reflects confidence for incorrect responses. Type 2 *D'* and type 2 *C* are defined as:

$$D' = Z(HR) - Z(FAR), \qquad C = \frac{-(Z(HR) + Z(FAR))}{2}$$

where Z is the standard Z-score, i.e. the inverse cumulative density function of the standard normal distribution. To distinguish type 2 variables from their type 1 counterparts we denote type 1 variables in lower-case (e.g. type 1 *d'*) and type 2 in upper-case (e.g. type 2 *D'*).

It is known that type 2 *D'* is highly biased by both type 1 and 2 thresholds (Barrett et al., 2013; Evans & Azzopardi, 2007; Galvin et al., 2003). An alternative measure is the 'bias-free' meta-*d'.* This is an estimate of the type 1 *d'* an SDT-optimal observer would need to have to generate the type 2 performance shown (for an in-depth explanation see Barrett et al., 2013 or Maniscalco & Lau, 2012). Importantly, meta-*d'* is measured in the same units as *d'*. This permits a direct comparison between objective and subjective sensitivity. Considering meta-*d'* as a proportion of *d*' gives us metacognitive efficiency, or the amount of type 1 information that is carried forward to the type 2 level. To take advantage of this feature we additionally analyzed our results using meta-*d'/d'.* We calculated meta-*d'*-balance from freely available online code (Barrett et al., 2013). This calculation was supplemented by a maximum likelihood estimation of SD_{noise}:$SD_{signal+noise}$ from the group-level data, also using freely available online code (columbia.edu/~bsm2015/type2sdt; Maniscalco & Lau, 2012).

As described in the introduction, we hypothesized that metacognitive performance would be improved when type 1 decisions are based on prior expectations. Testing this hypothesis requires comparing decisions which were based on (i.e. congruent with) prior expectations with those which were not. In the 25% condition, target absence is most probable meaning that an 'absent' report would be expectation-congruent and a 'present' report would be incongruent. The

opposite would be true for the 75% condition. We therefore computed, for each condition, type 2 *D'* following 'present' responses (hits and false alarms) and type 2 *D'* following 'absent' responses (misses and correct rejections). Analogous response-conditional meta-*d'* estimates were obtained from freely available online code (see Barrett et al., 2013, supplementary materials). Unfortunately, response-conditional meta-*d*′ is unlikely to be robust to criterion shifts like its response-unconditional counterpart (Barrett et al., 2013).

For all type 2 measures, a significant response by expectation interaction would demonstrate an effect of congruency. Note that we could not use a standard (i.e. response-unconditional) *D*′ or meta-*d*′ measure, because in this case degraded metacognition following one response could cancel out the improved metacognition following the alternative response.

3. Results

3.1 Expectation can be separated from attention. To verify that the concurrent visual search task successfully manipulated attention, we compared the contrast thresholds obtained in the full and diverted attention staircases. As expected, a one-tailed paired *t*-test revealed a significant increase in contrast in the dual-task (M = 0.080, SE = 0.011) relative to the single-task (M = 0.032, SE = 0.002) conditions, bootstrapped $t(18) = 4.64$, $p = .001$, 95% CI = [-0.06, -0.03], dz = 1.06. Thus, the paradigm successfully manipulated attention.

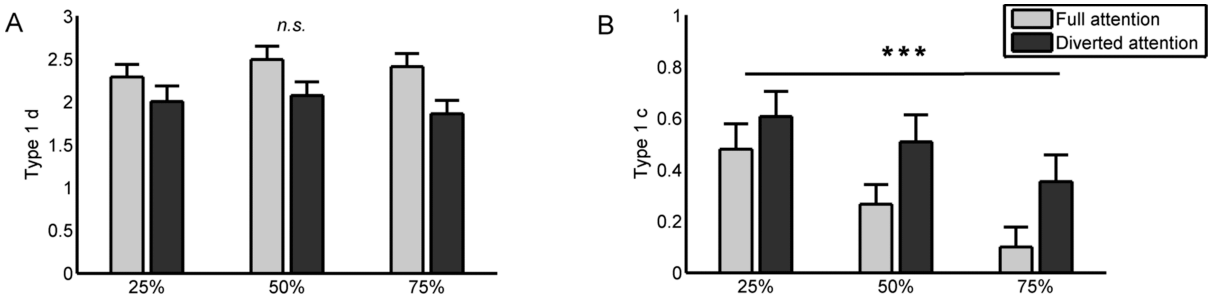

Figure 2. Type 1 results. Error bars are within-subjects SEM. **2*A***. Type 1 d' as a function of expectation and attention. **2*B*.** Type 1 criterion c as a function of expectation and attention. *** $p < .001$, ** $p < .01$, * $p < .05$, *n.s.* non-significant

Next, the effects of expectation and attention on each of (Gabor) detection sensitivity *d'* and (Gabor) decision threshold *c* were examined. These analyses addressed three questions: first, whether *d'* had been successfully equated across levels of attention and expectation; second, whether the expectation manipulation successfully biased *c*; third, whether expectation and attention were successfully separated at the type 1 level (i.e. did not interact under *d'* or *c*).

First, we performed a repeated-measures Expectation (0.25, 0.5, 0.75) x Attention (full, diverted) analysis of variance (ANOVA) on type 1 *d'*. This revealed that sensitivity did not significantly differ across the full ($M = 2.39$, $SE = 0.16$) and diverted ($M = 2.00$ $SE = 0.18$) attention conditions, $F(1, 18) = 3.03$, $p = .099$, $\eta_p^2 = .144$, or across Expectation conditions ($M_{.25} = 2.15$, $SE_{0.25} = 0.11$, $M_{.50} = 2.28$, $SE_{.50} = 0.15$, $M_{.75} = 2.14$, $M_{.75}$ 0.13), $F(2, 36) = 2.12$, $p = .124$, $\eta_p^2{}_E = .101$ (Figure 2A). Type 1 sensitivity was therefore successfully equated across all six conditions. This means that any changes in type 2 sensitivity cannot be attributed to changes in the amount of type 1 information. There was no significant interaction between Attention

and Expectation under *d', F(*2, 36) = 1.12, p = .34, η_p^2 = .059, suggesting that the two factors were successfully separated with respect to type 1 detection performance.

A repeated-measures Expectation (0.25, 0.5, 0.75) x Attention (full, diverted) analysis of variance (ANOVA) under decision threshold *c* revealed a significant main effect of Expectation, F(2, 36) = 9.18, p = .001, η_p^2 = .338. A trend analysis demonstrated that decision threshold linearly liberalized (more likely to report target present) as the probability of target presence increased, F(1, 18) = 15.72, p =.001, η_p^2 = .466. The paradigm therefore successfully manipulated expectation. Attention had no significant main effect on decision threshold, F(1,18) = 0.93, p =.93, η_p^2 = .148 and did not significantly interact with Expectation, F(2,36) = 0.85, p = .434, η_p^2 = .045 (Figure 2B) . Therefore attention and expectation were separated with respect to type 1 decision threshold, as well as type 1 sensitivity.

In the diverted attention condition, participants were instructed to perform the detection and the visual search task simultaneously, prioritizing visual search. However, if participants were unable to divide their attention across the two tasks then we would expect a significant negative correlation between trial-by-trial Gabor detection and visual search accuracy. To address this question we computed the Spearman's correlation coefficient between trial-by-trial detection accuracy scores on the two tasks for each participant. A one-sample bootstrapped *t*-test against zero revealed that at the group-level there was no significant trade-off in performance between the two tasks, M = 0.02, SD = 0.09, t(18) = 0.94, p = .361, 95% CI [-.023, .059]. Thus, participants were able to perform the two perceptual tasks simultaneously.

Participants were able to perform the tasks simultaneously, but if the visual search task interfered with Gabor detection sensitivity we might expect a significant negative correlation between experiment-wise performance in the two tasks. To address this concern we calculated *d'* and *c* for the visual search responses and correlated them with their Gabor detection counterparts. Across participants there was no significant (Pearson's) correlation between visual search *d'* and (diverted attention) Gabor *d' r*(19) = .250, *p* = .302, bootstrapped 95% CI [-.326, .623]. Similarly, there was no significant (Pearson's) correlation between type 1 decision thresholds for the two tasks, *r*(19) = .359, *p* = .131, bootstrapped 95% CI [-.043, .723]. These results suggest that performing the visual search task did not significantly interfere with performing the Gabor detection task. This, combined with the absence of a negative correlation between trial-by-trial accuracy on the two tasks and with the absence of attention by expectation interactions under *d'* and *c*, demonstrates that attention and expectation were sufficiently separated at the type 1 level.

The results so far indicate that the paradigm successfully influenced both expectation (participants were more likely to report target absence when the probability of target presentation was low than when it was high) and attention (contrast sensitivity was reduced when attention was diverted). Furthermore, they indicate that expectation and attention did not significantly interact. Given this, we were able to examine how metacognitive sensitivity is specifically affected by expectation and attention, without confounds of task difficulty.

3.2 Expectation improves metacognitive performance

Our main hypothesis was that metacognition would be improved following an expectation-congruent response. In the 25% condition, where target absence is expected, misses and correct rejections ('no') would be expectation-congruent responses and false alarms and hits ('yes') would be incongruent. The reverse is true for the 75% condition, where target presence is expected.

To test our hypothesis, response-conditional type 2 *D's* (see Methods) were subjected to a repeated-measures Expectation (0.25, 0.5, 0.75) x Attention (full, diverted) x Report (present, absent) analysis of variance (ANOVA).

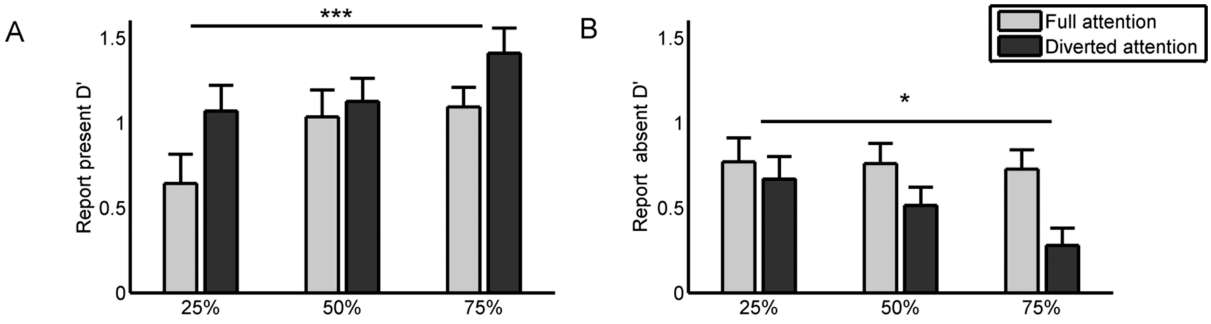

Figure 3. Response-conditional type 2 D' as function of expectation and attention. Black lines indicate linear changes in D' with expectation, independently of attention. **(A)** Type 2 D' for reports of target presence increases with expectation of presence **(B)** Type 2 D' for reports of target absence increases with expectation of absence**.** Error bars are with-subjects SEM. * $p < .05$ ** $p < .01$, *** $p < .001$.

Critically, the ANOVA revealed a significant two-way interaction between Expectation and Report, $F(2,36) = 5.60$, $p = .008$, $\eta_p^2 = .238$ (figure 3). To further probe this effect we collapsed across attention conditions and performed *a priori* trend analyses. *D'* for target present reports exhibited a significant linear trend with Expectation, $F(1,18) = 13.85$, $p = .001$ (1-tailed), $\eta^2 = .435$ such that as the probability of target presentation increased from 25% (target presence improbable)

to 75% (target presence probable), type 2 *D'* increased (Figure 3A). Similarly, when participants reported the Gabor as absent there was a significant linear trend with Expectation in the opposite direction, $F(1,18) = 3.83$, $p = .033$ (1-tailed), $\eta^2 = .175$: as the probability of target presentation decreased from 75% (target absence improbable) to 25% (target absence probable), type 2 D' increased (Figure 3B). This congruency effect supports our hypothesis that expectation improves metacognition.

As well as a significant Report x Expectation interaction, there was a significant interaction between Report and Attention, $F(1,18) = 5.61$, $p = .029$, $\eta_p^2 = .238$. This interaction was driven by the presence of a significant difference between *D'* for absent and present reports under diverted attention ($M = 0.49$, $SE = 0.13$ and $M = 1.20$, $SE = 0.19$, respectively) , $F(1,18) = 6.32$, $p = .022$, $\eta^2 = .260$, but not under full attention ($M = 0.75$, $SE = 0.11$ and $M = 0.92$, $SE = 0.16$, respectively), $F(1,18) = 0.84$, $p = .372$, $\eta^2 = .045$. This unexpected result suggests that inattention disrupts metacognition for unseen but not seen targets.

The ANOVA did not reveal a significant main effect of Expectation on *D',* $F(2,36) = 0.64$, $p = .533$, $\eta_p^2 = .034$. This is unsurprising, because the influence of expectation is seen by comparing expectation-congruence relative to incongruence. There was also no significant main effect of Attention on type 2 *D',* $F(1,18) = 0.01$, $p = .953$, $\eta_p^2 = .001$, and no significant Report by Attention by Expectation interaction, $F(1.60,28.81) = 0.11$, $p = .858$, $\eta_p^2 = .006$ ($\varepsilon = .748$, Huynh-Feldt corrected).

In summary, these data under type 2 *D'* indicate that metacognitive performance improved when reports of target absence or presence were congruent with participants' expectation (25% or 75% condition, respectively), as compared to when they were incongruent (75% or 25% condition respectively).

3.3 Expectation liberalizes confidence judgments

Given that expectation improved metacognitive performance, did expectations also increase subjective confidence? Type 2 confidence threshold can be interpreted as a proxy measure of the strength of the perceptual experience (Fleming & Lau, 2014). We therefore asked whether expectation-congruent reports were associated with higher confidence ratings than their incongruent counterparts. Such a result could be interpreted as expectations strengthening the associated perceptual experience.

We tested this possibility by asking whether expectation and report interacted under confidence threshold *C.* Confidence threshold is analogous to type1 decision threshold, signaling over-confidence when it is negative and under-confidence when it is positive. Therefore, if expectation liberalizes confidence judgments we would expect confidence thresholds for 'present' responses to liberalize with increased expectation of presence. Following an 'absent' response, we would expect confidence to liberalize with increasing expectation of target absence (i.e. *decreasing* expectation of target presence).

To test this possibility we ran a repeated-measures Expectation (0.25, 0.5, 0.75) x Attention (full, diverted) x Report (present, absent) analysis of variance (ANOVA) on *C*. This revealed a significant three-way interaction, $F(2,36) = 4.69$, $p = .015$, $\eta_p^2 = .207$, which was not found in the ANOVA on type 2 *D'*. We analyzed this interaction by performing simple effects analyses separately for the full and diverted attention conditions. Under full attention, Report and Expectation significantly interacted, $F(2,36) = 15.95$, $p < .001$, $\eta_p^2 = .470$. The pattern was as found for type 2 *D'*: with increasing probability of target presence, there was a linear decrease in type

2 *C* (more likely to report confidence) when the target was reported as present, $F(1,18) = 11.48$, $p = .002$, (one-tailed) $\eta^2 = .272$, and a linear increase in type 2 *C* (more likely to report guess) when the target was reported as absent, $F(1,18) = 25.29$, $p < .001$ (one-tailed), $\eta^2 = .584$. Thus, under full attention expectations liberalize subjective confidence judgments.

By contrast, under diverted attention there was neither a significant main effect of Expectation, $F(1,18) = .339$, $\eta_p^2 = .051$, nor a significant interaction between Expectation and Report, $F(2,36) = 2.84$, $p = .082$, $\eta_p^2 = .136$.

The ANOVA under *C*, revealed no significant main effect of Attention, $F(1,18) = 0.83$, $p = .374$, $\eta_p^2 = .044$, and no significant interactions between Attention and Report, $F(1,18) = 4.09$, $p = .058$, $\eta_p^2 = .185$, or Attention and Expectation $F(1,18) = 0.83$, $p = .444$, $\eta_p^2 = .044$.

Summarizing so far, metacognition improved for expectation-congruent perceptual decisions, independently of whether attention was focused on or diverted from the task. This effect was mirrored under confidence thresholds, but only under full attention. Therefore under conditions of full attention only, the perceptual experience associated with expectation-congruent decisions may be stronger than that for expectation-incongruent decisions.

3.4 Report-expectation congruency increases meta-d'.

To assure the robustness of our findings under type 2 *D'*, we re-analyzed the data using response-conditional meta-*d'*. As mentioned in section 2.5.2, given the type 2 performance observed, meta-*d'* is the type 1 *d'* that would be expected from the SDT-optimal observer who used all of the available type 1 information. Meta-*d'*/*d'*

is therefore the proportion of type 1 information used in the type 2 decision. We expected to find the same pattern of results as those obtained under *D'* – a Report by Expectation interaction whereby meta-*d'*/*d*' increases with response-expectation congruency. Only 1/19 of our participants fully met the criteria for assuring reliable meta-*d'* estimates (for all 6 conditions, 0.05 ≤ hr, far, HR_+, FAR_+, HR_-, FAR_- ≤ 0.95; see Barrett et al., 2013). We therefore retained participants who met these criteria in at least 3/6 conditions. This left us with 12 participants for the analysis.

As for the previous analyses, a repeated-measures Expectation (0.25, 0.5, 0.75) x Attention (full, diverted) x Report (present, absent) analysis of variance (ANOVA) was conducted, but this time using meta-*d'*/*d* as the dependent variable.

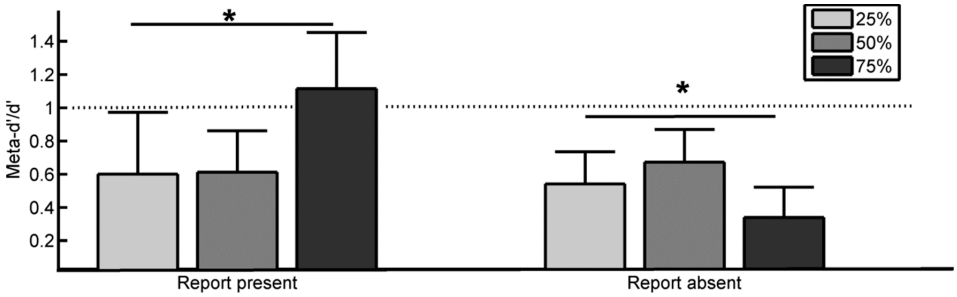

Figure 4. Meta-d'/d' as function of expectation and type 1 report. Error bars are with-subjects SEM. * $p < .05$, ** $p < .01$, *** $p < .001$.

Consistent with our previous result, the analysis revealed a significant Expectation x Report interaction, $F(2,22) = 8.75$, $p = .002$, $\eta_p^2 = .443$. *A priori* trend analyses revealed that following a 'present' response, meta-*d'*/*d'* linearly increased with expectation of target presence, $F(1,11) = 5.12$, $p = .022$ (one-tailed), $\eta^2 = .318$. Following an 'absent' response there was a significant decrease in meta-*d'*/*d'* as the probability of target presence increased, $F(1,11) = 4.22$, $p = .032$ (one-tailed), $\eta^2 = .277$. These patterns are illustrated in figure 4. We found no other significant main (all $F < 2.37$, all $p > .15$, all $\eta_p^2 < .29$) or interaction (all $F < 0.99$, all $p > .32$, all $\eta_p^2 <$

.09) effects. This pattern of results held under slightly narrower and broader exclusion criteria (i.e. proportion of stable conditions).

Summarizing, report-expectation congruency improves metacognitive performance when measured by response-conditional meta-*d'*, as well as when measured by response-conditional *D'*.

3.5 A Type 2, Bayesian Signal Detection Theoretic Model of Expectation and Top-Down Attention

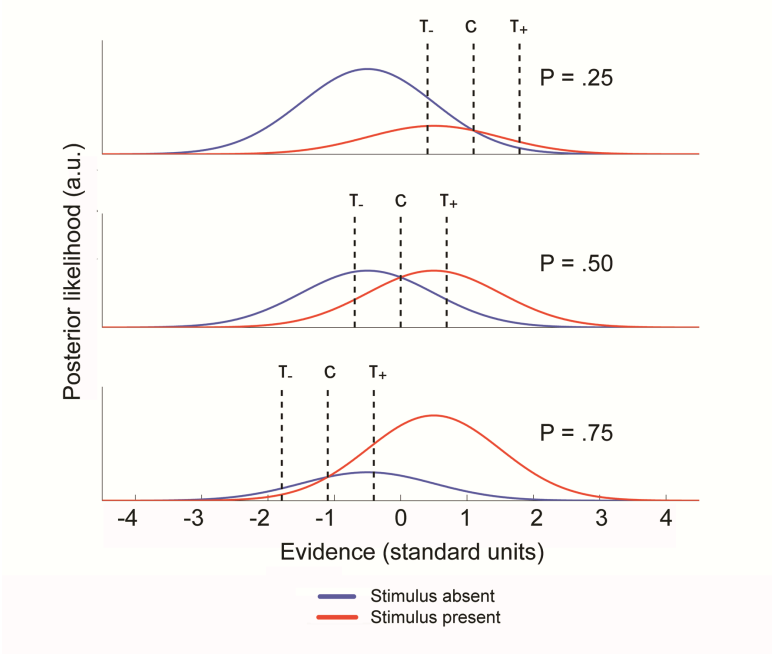

Figure 5. A Bayesian signal detection theoretic model of prior expectation. Each panel plots the posterior likelihood of a perceptual event against the evidence given distinct prior probabilities (p) of stimulus present. The blue curve represents the event of stimulus absence and the red curve, stimulus presence. Type 1 d' (the distance between the blue and red Gaussians) is held at 1. The curves are aligned so that criterion is unbiased when p = .50. The dashed lines show the decision (c) and confidence (T_+, T_-) thresholds. These are each determined by a fixed posterior likelihood ratio R for stimulus present to stimulus

absent. These plots illustrate that detection, as well as confidence about detection, liberalizes with increased prior expectation on Bayesian SDT.

To model the influence of top-down expectation on metacognitive sensitivity we extended standard signal detection theory (SDT) to incorporate prior expectations (Figure 5). In our model, the evidence is the internal variable '*x*' in SDT (the internal representation of Gabor contrast) and the expectation is the probability of Gabor patch presentation. The 'signal' and 'noise' distributions were reformulated as posterior distributions of the cases of target present and absent, given both the evidence and the expectation. Type 1 and 2 decision criteria (*c* and *C*) were formulated as distinct thresholds for the posterior ratio of probabilities of present (S=1) to absent (S=0). For probability *p* of stimulus present and evidence *x*, this ratio, which we denote by R, is given by

$$R = \frac{P(S=1|x)}{P(S=0|x)} = \frac{P(x|S=1)P(S=1)}{P(x|S=0)P(S=0)} = \frac{\varphi_{d',\sigma}(x) \times p}{\varphi_{0,1}(x) \times (1-p)}$$

where $\varphi_{\mu,s}$ is the probability density function of a normal distribution with mean μ and standard deviation *s*. Assuming the SDT model, this ratio monotonically increases with the evidence *x*. To model the effect of diverted attention we implemented the solution proposed by Rahnev et al. (2011), in which inattention increases the trial-by-trial internal noise. To assess whether this model could account for our data we computed the response-conditional type 2 *D'*s predicted by the model at varying, continuous levels of prior expectation of patch present. This was done separately for the full and diverted attention cases.

Parameters were determined in the following way: Type 1 *d'* was set to 2.39 and 2.00 for the full and diverted attention conditions respectively, reflecting the mean empirical values we obtained. For each level of attention, the type 1 and 2 thresholds for R were based on the mean empirical type 1 and 2 hit and false alarm rates in the respective 50% expectation condition. For the full attention case, the obtained type 1 threshold was R=1.88, and the upper and lower type 2 thresholds were R=4.27 and R=0.68 respectively. For the diverted attention case, these were respectively R=2.52, R=4.06 and R=0.86. For full details on obtaining type 1 and 2 decision thresholds from type 1 and 2 hit and false alarm rates, see Barrett et al. (2013). Notice that, since contrast was increased in the experiment for diverted attention, the models for full and diverted attention were approximately the same; only the threshold values (R) differed slightly.

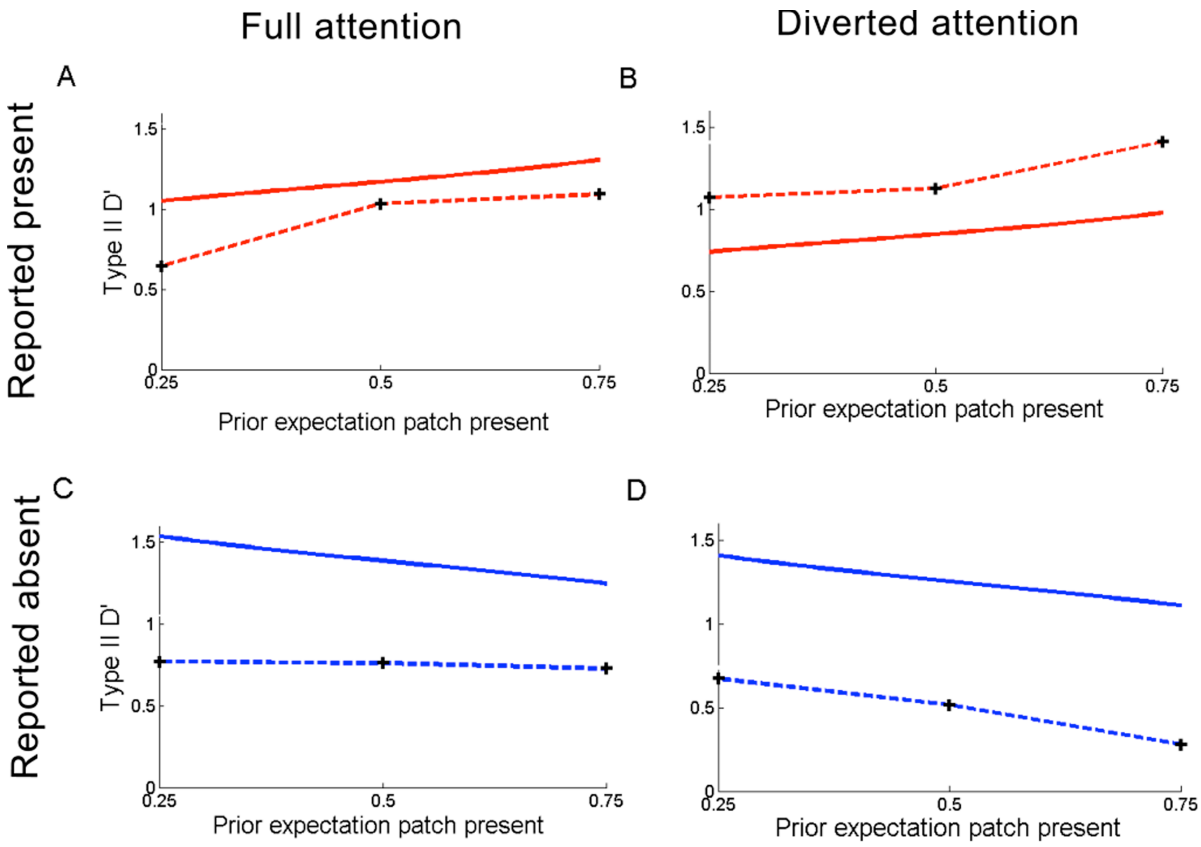

Figure 6. Modelling of empirical results. Solid lines represent stimulated results over continuous probabilities of target present. Dashed lines are the - corresponding empirical results collected over 25, 50 and 75% of target presence. The top and bottom rows show results for reported present and absent trials, respectively. The leftward and rightward columns show results for full and diverted attention conditions, respectively.

Figures 6A-D compare the predicted and empirical *D'*s across levels of report and attention. In agreement with the empirical data, predicted *D'* for positive responses increased with prior expectation for target present (Figures 6A and 6B), while *D*' for negative responses, it decreased (Figures 6C and 6D). As was the case for the empirical results, this decrease demonstrates an increase in *D*' with increased prior expectation for target absent. The model predicted slight attentional

modulations of *D*', which reflect numerical differences in empirical type 1 *d*'. Simulated *D*' values for 'absent' responses also took substantially higher values than those collected empirically. Moreover, simulated *D'* was higher for absent than for present responses, whereas the reverse trend was found empirically. These two features persisted for variant models on which signal and noise distribution variances were unequal. They are likely attributable to asymmetries in the degradation of type 1 evidence available for metacognition, an investigation of which is beyond the present scope.

In summary, our modeling analyses demonstrate that the observed dependencies of metacognitive performance on prior expectation are consistent with a signal-detection theory model extended according to Bayesian principles to incorporate expectations as priors.

3.6 Effect of expectations on concurrent visual search task

So far we have shown that expectations of Gabor presence or absence improve metacognition for the Gabor detection task. Given this, could expectations of Gabor target presence or absence also facilitate perceptual decisions for the visual search task? The expectations induced by the paradigm pertained to the Gabor target, however the influence of these expectations may free perceptual and cognitive resources for other tasks.

To address this question, we first asked whether expectation affects decisions made on the visual search task (i.e. T presence or absence). This was achieved by computing type 1 *c* for the visual search task as a function of expectation. Visual

search data from the full attention condition could not be analyzed because the required responses were not collected.

A one-way Expectation (.25, .50, .75) repeated measures ANOVA under visual search criterion c_{vs} revealed a significant effect of Expectation, $F(2,36) = 6.17$, $p = .005$, $\eta^2 = .255$. However, rather than expectation of Gabor presence inducing a liberal criterion shift under the visual search task, as it did under the Gabor task, there was a significant quadratic trend, $F(1,18) = 11.74$, $p = .003$, $\eta^2 = .395$. This trend was such that participants were more likely to report that a T was present (liberal shift) in the 50% condition ($M = 0.19$, $SE = 0.09$) than when they had a task-irrelevant prior expectation of Gabor presence or absence (25% and 75% conditions, $M = 0.35$, $SE = 0.09$, $M = 0.32$, $SE = 0.08$, respectively). Therefore the task-irrelevant expectation of Gabor presence or absence did not bias participants towards reporting presence or absence on the visual search task. Rather, expectations induced a conservative shift in *c* relative to the neutral (50%) condition.

Given that expectation of Gabor presence or absence biased decisions made in the visual search task, they may also have affected sensitivity. We therefore calculated visual search *d'* as a function of Gabor detection accuracy and expectation-Gabor response congruence. The factor Congruence was formed by grouping trials according to whether the response to the Gabor task (present or absent) was congruent or incongruent with the prior expectation (75%, where they expect presence, 50%, which is neutral, 25%, where they expect absence). This factor therefore represents the influence of expectation on Gabor decision. If visual search performance is modulated by the effect of expectation on the Gabor task then there should be an effect of this factor.

A repeated-measures Gabor accuracy (correct, incorrect) x Gabor congruence (incongruent, neutral, congruent) ANOVA on visual search d'_{vs} revealed a significant main effect of Gabor accuracy, $F(1,18) = 4.80$, $p = .015$, $\eta_p^2 = .288$, whereby d'_{vs} was higher following a correct ($M = 1.72$, $SE = 0.16$) than an incorrect ($M = 1.31$, $SE = 0.18$) response on the Gabor detection task. Therefore high perceptual sensitivity for the Gabor was associated with high perceptual sensitivity for the visual search task as well. The ANOVA also revealed a marginally significant interaction between accuracy and congruence, $F(2,36) = 2.95$, $p = .065$, $\eta_p^2 = .141$. Post-hoc trend analyses revealed that d'_{vs} linearly increased with expectation-Gabor response congruence following a correct response on the Gabor task, $F(1,18) = 4.49$, $p = .048$, $\eta^2 = .200$ and linearly decreased with congruency following an incorrect Gabor response, $F(1,18) = 5.27$, $p = .034$, $\eta^2 = .226$. This result suggests that visual search sensitivity improved when the (Gabor) expectation had been valid (i.e. met in the stimulus-conditional sense). This follows from the observation that the expectation was only valid in trials where correct and congruent or incorrect and incongruent responses were made. To illustrate, in the 25% condition, correct responses were correct rejections (congruent, valid expectation) or hits (incongruent, invalid expectation). The former was associated with a higher d'_{vs} than the latter. Incorrect responses were misses (congruent, invalid expectation) or false alarms (incongruent, valid expectation). Here, the latter was associated with a higher d'_{vs} than the former. Thus perceptual sensitivity for the attended task was facilitated by valid (task-irrelevant) expectations for the unattended task.

Discussion

In this paper we have shown that the facilitatory effects of prior expectation on perceptual decision also manifest their influence in metacognitive judgments. We developed a target detection paradigm in which the probability of target presence was manipulated block-wise. This probability, of which participants were informed, significantly biased decision thresholds in the expectation-congruent direction, while leaving sensitivity *d'* unaffected (as ensured by our staircase procedure). In this way we avoided confounding increased type 2 sensitivity with increased type 1 sensitivity (Lau & Passingham, 2006), and were able to assess metacognition, indexed by both type 2 *D'* and meta-*d'*, as a function of perceptual decision and prior expectation. Our main finding was that metacognitive sensitivity increased for expectation-congruent as compared to expectation-incongruent perceptual decisions. Metacognitive sensitivity is determined according to the trial-by-trial correspondence between confidence and accuracy. Importantly, because we offered no trial-by-trial information about the probability of target occurrence, our results cannot be attributed to a trivial relationship between an expectancy cue and the subsequent report. Rather, we found a shift in type 1 threshold with expectation, and a liberalization of type 2 threshold following an expectation-congruent response to an attended target. This suggests that basing decision on prior expectations induces a superior placement of type 1 and 2 thresholds for metacognition.

Our effect of expectation on confidence required attention, consistent with some previous work in type 1 tasks (Chennu et al., 2013; Hsu et al., 2014; Jiang et al., 2013; but c.f. Kok, Jehee, & de Lange, 2012). However, analyses using both type 2 *D'* and meta-*d'* revealed that expectation improved metacognition independently of attention. We also found no significant difference in metacognition for perceptual decisions made under full attention relative to those made under diverted attention

(though under diverted attention, metacognition differed as a function of report). Though perhaps counter-intuitive, this invariance of metacognition to attention is consistent with recent work showing that metacognition is preserved for visual sensory memory, which does not require attention (Vandenbroucke et al., 2014). It is also consistent with research demonstrating above-chance metacognitive sensitivity for unattended and unseen target stimuli (Kanai et al., 2010).

Measuring metacognition

To assess how metacognition is affected by expectation we used the type 2 signal detection theory (SDT) measure *D′*. However, the type 2 SDT model underlying *D'* assumes that the probability of making a correct or an incorrect response can be modeled as Gaussian distributions over a type 2 decision axis. This formulation is problematic because such distributions are usually impossible to achieve (Evans & Azzopardi, 2007; Galvin et al., 2003). This issue means that *D′* will not be invariant to type 1 or type 2 criterion shifts (Barrett et al., 2013; Evans & Azzopardi, 2007). In the present study, expectation induced both type 1 and type 2 criterion shifts. As a result, we cannot distinguish between two possible reasons for why *D′* may have increased for expectation-congruent responses. One possibility is that expectation increased the quantity of information available for the type 2 judgment (metacognitive efficacy, Fleming & Lau, 2014). Alternatively, the increase in *D′* could have been driven by a change in criteria placement that indirectly optimized metacognitive sensitivity. Importantly, under both scenarios metacognitive performance, as measured by *D′*, nevertheless improved. The liberalization of confidence threshold by expectation, though a source of bias in the numerical value *D′* will take, can be interpreted as reflecting the strength of the perceptual experience

(Fleming & Lau, 2014). Therefore rather than being unequivocally problematic, type 2 criteria shifts speak to subjective components of perception.

Our finding that expectation increased *D'* was replicated using the measure meta-*d'* (see section 2.5.2. Barrett et al., 2013; Maniscalco & Lau, 2012). Meta-*d'* is robust to changes in type 1 and 2 criteria, however response-conditional meta-*d'* – as required by the analyses presented in this paper - is not (Barrett et al., 2013). The invariance is lost because meta-*d'* measures remove bias by taking a weighted average of the (biased) response-conditional measures. Therefore while we replicated our effect using meta-*d',* we are still unable to ascertain whether expectation improves metacognitive *efficacy* or not. Nevertheless, our results under type 2 *D'* and meta-*d'* together provide converging evidence for the facilitatory effect of expectation on metacognition.

Modeling metacognition

The framework of SDT applied to visual perception emphasizes the importance of 'bottom-up' processing, whereby afferent sensory signals are repeatedly transformed to generate perceptual decisions at both objective (type 1) and subjective (type 2) levels. However, our data add to an increasing body of work which has demonstrated the importance of top-down processes in shaping perceptual decisions (Bar et al., 2006; Gilbert & Li, 2013; Wacongne et al., 2011). Together, these data pose a challenge to bottom-up models of perception and are difficult to reconcile with standard expressions of SDT.

To formally account for these top-down effects within SDT, we developed a type 2 Bayesian signal detection model which models prior expectations by defining

decision threshold as the posterior odds of a target being present. This model successfully predicted an increase in type 2 *D'* following expectation-congruent responses. Diverted attention was modeled by increasing internal noise - as recently proposed by Rahnev et al. (2011). This successfully predicted that the influence of expectation on *D'* would be independent of attention.

We recognize that our model did not capture all aspects of the observed data. In particular, the model predicted an improvement in metacognition following a target absent response, but this was not found empirically. This discrepancy is likely to have arisen from influences on metacognition which were not included in our model, such as the incorporation of additional sources of information relevant to perceptual decision. Nonetheless, by accounting for the main effects of (top-down) prior expectations on *D',* we have demonstrated the scope for formal synthesis between the traditionally 'bottom-up' signal detection theory and 'top-down' influences characteristic of alternative frameworks like 'predictive coding' or the Bayesian brain (Beck et al., 2009; Clark, 2013; Friston, 2009; J. Hohwy, 2013; Lee & Mumford, 2003).

From SDT to the Bayesian brain

The increasingly influential predictive coding framework views the brain as a Bayesian hypothesis-tester, and explains perceptual decision as an inference about the most likely cause of incoming sensory input (Clark, 2013; Rao & Ballard, 1999; Seth, 2014). In this view, top-down expectations constrain perceptual decision according to the prior likelihood of that decision. The sensory input remaining unexplained is termed prediction error, and only this percolates upwards in the sensory hierarchy (Friston, 2010; Rao & Ballard, 2004; Spratling, 2008). The

eventual perceptual choice will be the perceptual hypothesis with the highest posterior probability. This framework fits comfortably with our novel finding that under dual-task conditions, sensitivity for the attended (visual search) task was increased when participants held valid expectations pertaining to the unattended (Gabor) task: when prior expectations facilitate decision for the unattended Gabor task, additional processing resources should be available for the attended visual search task (Hohwy, 2012).

Certain predictive coding formulations also explicitly model the importance of the reliability (or 'precision') of the bottom-up signal to perception (e.g. Feldman & Friston, 2010). In this paper we have shown that expectations liberalized subjective confidence judgments for attended (i.e. high precision) targets. Previous work has shown that confidence judgments are a function of both sensory evidence and internal noise (Kepecs, Uchida, Zariwala, & Mainen, 2008; Yeung & Summerfield, 2012; Zylberberg, Barttfeld, & Sigman, 2012; Zylberberg, Roelfsema, & Sigman, 2014). This relationship has been likened to a *p*-value, which quantifies the evidence for a hypothesis (mean) and scales with the reliability of that evidence (standard error; Kepecs & Mainen, 2012). In fact, such a formulation of confidence is highly compatible with predictive coding. Bringing these together, decisional confidence could be explained in predictive coding terms, where the mean is the posterior probability of a perceptual hypothesis, and the standard error is the precision of the evidence (Feldman & Friston, 2010). Such a conceptualization of confidence would explain the congruency-attention interaction found in this paper. It is also consistent with work demonstrating that confidence evolves together with the decision variable (De Martino, Fleming, Garrett, & Dolan, 2013; Fetsch, Kiani, Newsome, & Shadlen, 2014; Kepecs et al., 2008).

The above account may explain the construction of confidence judgments within a single level of the perceptual hierarchy. However, successful metacognitive evaluations and the subjective aspect of decisional confidence may be a function of uncertainty estimates over multiple hierarchical levels. We leave the theoretical and neural underpinnings of how expectation modulates metacognition open to future research.

Conclusions

In summary, we show for the first time that top-down prior expectations can influence metacognition for perceptual decision, illustrating the action of priors on complex cognitive functions. Specifically, we found that perceptual decisions which are congruent with valid perceptual expectations lead to increased metacognitive sensitivity, independently of attentional allocation. This finding motivated the development of an extended Bayesian signal detection theoretic model which incorporates top-down prior expectations, and moreover, formally integrates two previously distinct frameworks for perceptual decision: (top-down) predictive coding and (bottom-up) signal detection theory. Finally, measures of metacognition are often used as an indirect measure of awareness (Kanai et al., 2010; Kunimoto et al., 2001; Seth, Dienes, Cleeremans, Overgaard, & Pessoa, 2008). Therefore, by demonstrating increased metacognitive sensitivity for expected perceptual events, we provide evidence for the existence of a mechanism, based on prior expectations, that underpins metacognitive sensitivity and contributes to our understanding of the brain basis of visual awareness.

Author contributions: M. T. Sherman, R. Kanai and A. K. Seth conceptualized and designed the study. M. T. Sherman collected and analyzed the data. A. B. Barrett performed the modeling analyses. M. T. Sherman drafted the manuscript and A. K. Seth, A. B. Barrett and R. Kanai provided critical revisions. All authors approved the final version of the manuscript for submission. This research was supported by the Dr. Mortimer and Theresa Sackler Foundation, which supports the Sackler Centre for Consciousness Science. A.B.B. is supported by Engineering and Physical Sciences Research Council Grant EP/L005131/1.

5. References

Banjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (methodological)*, *57*(1), 289–300.

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., … Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences*, *103*(2), 449–454. doi:10.1073/pnas.0507062103

Barrett, A. B., Dienes, Z., & Seth, A. K. (2013). Measures of Metacognition on Signal-Detection Theoretic Models. *Psychological Methods*. doi:10.1037/a0033268

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Shadlen, M. N., … Pouget, A. (2009). Probabilistic population codes for Bayesian decision making, *60*(6), 1142–1152. doi:10.1016/j.neuron.2008.09.021.Probabilistic

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

Chennu, S., Noreika, V., Gueorguiev, D., Blenkmann, A., Kochen, S., Ibáñez, A., … Bekinschtein, T. a. (2013). Expectation and attention in hierarchical auditory prediction. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *33*(27), 11194–205. doi:10.1523/JNEUROSCI.0114-13.2013

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences*, *36*(3), 181–204. doi:10.1017/S0140525X12000477

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105–10. doi:10.1038/nn.3279

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 193–222.

Duncan, J. (2006). EPS Mid-Career Award 2004: brain mechanisms of attention. *The Quarterly Journal of Experimental Psychology (2006)*, *59*(1), 2–27. doi:10.1080/17470210500260674

Egner, T., Monti, J. M., & Summerfield, C. (2010). Expectation and surprise determine neural population responses in the ventral visual stream. *Journal of Neuroscience*, *30*(49), 16601–16608. doi:10.1523/JNEUROSCI.2770-10.2010

Evans, S., & Azzopardi, P. (2007). Evaluation of a “bias-free” measure of awareness. *Spatial Vision*, *20*(1-2), 61–77. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17357716

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4*(December), 215. doi:10.3389/fnhum.2010.00215

Fetsch, C. R., Kiani, R., Newsome, W. T., & Shadlen, M. N. (2014). Effects of Cortical Microstimulation on Confidence in a Perceptual Decision. *Neuron*, *83*(4), 797–804. doi:10.1016/j.neuron.2014.07.011

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*(July), 1–9. doi:10.3389/fnhum.2014.00443

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301. doi:10.1016/j.tics.2009.04.005

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews. Neuroscience*, *11*(2), 127–38. doi:10.1038/nrn2787

Friston, K., Adams, R. A, Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: saccades as experiments. *Frontiers in Psychology*, *3*(May), 151. doi:10.3389/fpsyg.2012.00151

Galvin, S. J., Podd, J. V, Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review*, *10*(4), 843–76. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15000533

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience*, *14*(May), 350–363. doi:10.1038/nrn3476

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1.). New York: Wiley.

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, *3*(April), 96. doi:10.3389/fpsyg.2012.00096

Hohwy, J. (2013). *The Predictive Mind*. Oxford: OUP.

Hsu, Y.-F., Hämäläinen, J. a, & Waszak, F. (2014). Both attention and prediction are necessary for adaptive neuronal tuning in sensory processing. *Frontiers in Human Neuroscience*, *8*(March), 152. doi:10.3389/fnhum.2014.00152

Jiang, J., Summerfield, C., & Egner, T. (2013). Attention sharpens the distinction between expected and unexpected percepts in the visual brain. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *33*(47), 18438–47. doi:10.1523/JNEUROSCI.3308-13.2013

Kanai, R., Walsh, V., & Tseng, C. (2010). Subjective discriminability of invisibility: a framework for distinguishing perceptual and attentional failures of awareness. *Consciousness and Cognition*, *19*(4), 1045–1057. doi:10.1016/j.concog.2010.06.003

Kepecs, A. & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1594), 1322–37. doi:10.1098/rstb.2012.0037

Kepecs, A., Uchida, N., Zariwala, H. a, & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*(7210), 227–31. doi:10.1038/nature07200

Kleiner, M., D., B., & Pelli, D. (2007). What's new in Psychtoolbox-3? In *Perception 36 ECVP Abstract Supplement*.

Kok, P., Jehee, J. F., & de Lange, F. P. (2012). Less Is More : Expectation Sharpens Representations in the Primary Visual Cortex. *Neuron*, *75*(2), 265–270.

Kok, P., Rahnev, D., Jehee, J. F. M., Lau, H. C., & de Lange, F. P. (2012). Attention reverses the effect of prediction in silencing sensory signals. *Cerebral Cortex (New York, N.Y. : 1991)*, *22*(9), 2197–206. doi:10.1093/cercor/bhr310

Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition*, *10*, 294–340.

Lau, H. C., & Passingham, R. E. (2006). Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 18763–18768. doi:10.1073/pnas.0607716103

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, *20*(7), 1434. doi:10.1364/JOSAA.20.001434

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology Press.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–30. doi:10.1016/j.concog.2011.09.021

Melloni, L., Schwiedrzik, C. M., Müller, N., Rodriguez, E., & Singer, W. (2011). Expectations change the signatures and timing of electrophysiological correlates of perceptual awareness. *Journal of Neuroscience*, *31*(4), 1386–1396. doi:10.1523/JNEUROSCI.4570-10.2011

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3–25.

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*(12), 1513–5. doi:10.1038/nn.2948

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10195184

Rao, R. P. N., & Ballard, D. H. (2004). Probabilistic Models of Attention based on Iconic Representations and Predictive Coding, (July).

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*(3), 165–175. doi:10.1080/17588921003632529

Seth, A. K. (2014). A predictive processing theory of sensorymotor contingencies: Explaining the puzzle of perceptual presence and absence in synaesthesia. *Cognitive Neuroscience*.

Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M., & Pessoa, L. (2008). Measuring consciousness: relating behavioural and neurophysiological approaches. *Trends in Cognitive Sciences*, *12*(8), 314–21. doi:10.1016/j.tics.2008.04.008

Spratling, M. W. (2008). Predictive coding as a model of biased competition in visual attention. *Vision Research*, *48*(12), 1391–408. doi:10.1016/j.visres.2008.03.009

Stefanics, G., Hangya, B., Hernádi, I., Winkler, I., Lakatos, P., & Ulbert, I. (2010). Phase entrainment of human delta oscillations can mediate the effects of expectation on reaction speed. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *30*(41), 13578–85. doi:10.1523/JNEUROSCI.0703-10.2010

Sterzer, P., Frith, C., & Petrovic, P. (2008). Believing is seeing : expectations alter visual awareness Neural basis for unique hues. *Current Biology*, *18*(16), R697–R698.

Summerfield, C., & de Lange, F. P. (2014). Expectation in perceptual decision making: neural and computational mechanisms. *Nature Reviews Neuroscience*, (October). doi:10.1038/nrn3838

Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences*, *13*(9), 403–409. doi:10.1016/j.tics.2009.06.003

Vandenbroucke, A. R. E., Sligte, I. G., Barrett, A. B., Seth, A. K., Fahrenfort, J. J., & Lamme, V. A. F. (2014). Accurate metacognition for visual sensory memory representations. *Psychological Science*. doi:10.1177/0956797613516146

Wacongne, C., Labyt, E., Van Wassenhove, V., Bekinschtein, T., Naccache, L., & Dehaene, S. (2011). Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proceedings of the National Academy of Sciences*, *108*(51), 1–6. doi:10.1073/pnas.1117807108

Wilimzig, C., & Fahle, M. (2008). Spatial attention increases performance but not subjective confidence in a discrimination task. *Journal of Vision*, *8*(5), 1–10. doi:10.1167/8.5.7.Introduction

Wyart, V., Nobre, A. C., & Summerfield, C. (2012). Dissociable prior influences of signal probability and relevance on visual contrast sensitivity. *Proceedings of the National Academy of Sciences*, *109*(16), 6354–6354. doi:10.1073/pnas.1204601109

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*(1594), 1310–21. doi:10.1098/rstb.2011.0416

Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience*, *6*(September), 79. doi:10.3389/fnint.2012.00079

Zylberberg, A., Roelfsema, P. R., & Sigman, M. (2014). Variance misperception explains illusions of confidence in simple perceptual decisions. *Consciousness and Cognition*, *27C*, 246–253. doi:10.1016/j.concog.2014.05.012

Figure Captions