Brain Multiphysics 2 (2021) 100030

Contents lists available at ScienceDirect

Brain Multiphysics

journal homepage: www.sciencedirect.com/journal/brain-multiphysics

Awareness and confidence in perceptual decision-making

Joshua Skewes [a,b,*], Chris Frith [d], Morten Overgaard [c,e]

[a] Department for Linguistics, Cognitive Science, and Semiotics, Aarhus University, Jens Chr. Skous Vej, 4, Building 1485, 8000 Aarhus C, Denmark
[b] Interacting Minds Centre, Aarhus University, Denmark
[c] Aarhus Institute of Advanced Studies, Aarhus University, Denmark
[d] Welcome Trust Centre for Neuroimaging, University College London, United Kingdom
[e] Centre for Functionally Integrative Neuroscience, Aarhus University, Denmark

ARTICLE INFO

Keywords:
Perceptual decision-making
Awareness
Confidence

ABSTRACT

Perceptual decision-making employs a range of higher order metacognitive processes. Two of the most important of these are perceptual awareness; or the clarity with which one reports seeing a perceptual stimulus, and response confidence; or the certainty one has about the correctness of one's own perceptual categorizations. We used a novel false feedback paradigm to investigate the relationships between these two processes. We asked people to perform a standard psychophysical detection task. We used feedback to selectively intervene either on our participants' trust in their own perceptual awareness of the stimulus, or on their confidence in their own responses. We measured the effects of these interventions on response accuracy; on reports of perceptual awareness; and on response confidence. We found that by undermining people's trust in their awareness of the sensory stimulus, we could reliably reduce their accuracy on the task. We suggest that the reason this occurred is that people came to rely less on evidence from their senses when making perceptual decisions. We conclude by suggesting that there is a not a one-to-one mapping between content in conscious experience and how that content is used in perceptual decision making, and that one's perception of the reliability of content also plays a role.

STATEMENT OF SIGNIFICANCE

This paper explores how different kinds of metacognitive state are related to one another and to perceptual decision making. Our focus is on the states of metacognitive confidence and perceptual awareness. We examine how an intervention on the reliability of these states influences performance in a perceptual detection task. We also examine how the intervention influences reports of the states themselves. The intervention we use is false feedback. For one group of participants, we tell them their perceptual judgement is wrong whenever they report they are uncertain in their choice (confidence intervention). For another group, we tell them their judgement is wrong whenever they report that their experience of the stimulus is unclear (awareness intervention). We find that both interventions reduce the accuracy of people's judgements, but that the awareness intervention is more effective. Also, we find that only the awareness intervention reduces reports of both metacognitive confidence in the response, and awareness of the stimuli. The confidence intervention does not influence either metacognitive state. These results suggest that we should understand confidence and awareness as separate higher level cognitive states, and that we should understand awareness as having a stronger causal role than confidence in perception and performance.

Introduction

It has become increasingly common to include measures of metacognition in studies of perceptual decision-making. The two most common measures are metacognitive confidence ratings, which are designed to measure subjective certainty in the correctness of one's perceptual choices [11], and the perceptual awareness scale (PAS), which is designed to measure the clarity with which one experiences the sensory stimulus [16,19]. These measures are often treated as interchangeable. However, it has been shown that the two kinds of measures yield

* Corresponding author at: Department for Linguistics, Cognitive Science, and Semiotics, Aarhus University, Jens Chr. Skous Vej, 4, Building 1485, 8000 Aarhus C, Denmark.
E-mail address: filjcs@cas.au.dk (J. Skewes).

https://doi.org/10.1016/j.brain.2021.100030
Received 17 February 2021; Received in revised form 29 June 2021; Accepted 29 June 2021
Available online 3 July 2021

different results, even when they are included in the same task for the same participants [13,15,17,18,23].

Metacognition is typically considered a "higher order state", in contrast to "first order" cognitive states. Mental states of the first order denote our perceptions, thoughts, and emotions that by most accounts can be conscious or unconscious [7]. Mental states of a second order are most often referred to as metacognitive, i.e. mental states that are about first order mental states [5,10].

Here we present a simple experiment designed to further investigate the relationship between the two measures. In this experiment, we asked participants to perform a visual discrimination task. On each trial, we also asked participants to report their sensory awareness of the stimulus they had just seen, and their metacognitive confidence in their response. As an experimental manipulation, we selectively intervened on participants' perceived reliability of the two types of metacognition. In one group (i.e. the false confidence rating group), whenever participants reported that they were "slightly confident" of their response, we told them it had been wrong, even if it was correct. In another group (i.e. the false perceptual awareness scale group), whenever participants reported that they only saw a "vague glimpse" of the stimulus, we told them their response had been wrong, even if it was correct. We then compared how these two interventions influenced participants' confidence ratings, perceptual awareness ratings, and choice accuracy in the task.

The results of the experiment suggest that such an intervention on the reliability of perceptual awareness can reduce peoples' awareness ratings, their confidence ratings, and even their choice accuracy. Surprisingly, we find that the same kind of intervention on metacognitive confidence does not have the same effects. When people are consistently told that any response about which they are only slightly confident is also wrong, then we find that only their choice accuracy is reduced. Moreover, the evidence for this effect is weaker than for the comparable effect of false feedback about perceptual awareness.

Method

Participants

96 Aarhus University students participated. Data from six participants were excluded because the experiment was not completed. All participants provided written informed consent, and the experiment was conducted in accordance with regional ethical guidelines. Because the paradigm was designed to specifically test general perceptual processes, to maximize anonymity at the point of data collection, no demographic data were collected.

Stimuli

Participants were presented with Gabor patches composed of sine-wave gratings with a spatial frequency of 0.63 cycles/mm, wrapped in a Gaussian envelope of mean 0.5 mm and standard deviation 2.12 mm.

The stimulus image consisted of two Gabor patches presented on opposite corners of an invisible square (Fig. 1). The square measured 63.5 mm x 63.5 mm, and was centred on a fixation cross at the middle of the screen. For each stimulus image, there was always one Gabor patch that was oriented vertically, and one patch that was tilted at an orientation between 0 and 45°. In each trial, the tilted patch could appear on either the left or the right side of fixation. The stimuli were post-masked with plaids made of two overlapping Gabor patches oriented at 45 and 315°.

The trial structure is presented in Fig. 1. A fixation cross was presented for a random inter-trial interval of 200–500 ms, followed by the stimulus image for 32 ms, followed by a fixation cross for a variable stimulus onset asynchrony (SOA), followed by the post-mask for 600 ms. Participants were then presented with a response screen prompting them to indicate which side of fixation the tilted patch had been presented in the stimulus image.

Stimuli were presented on a 22″ LED display with a resolution of 1920 × 1080 pixels. Participants viewed the stimuli at about 500 mm, and responded using the arrow keys on a standard keyboard. PAS and CR responses were recorded using the 1–4 number keys.

Tasks and feedback conditions

Before the experiment, participants were instructed on the distinction between perceptual awareness and response confidence. Separate rating scales were presented as measuring these two phenomena. To measure perceptual awareness, participants were presented with the perceptual awareness (PAS) scale [16], a four-point scale which quantifies the clarity of awareness of a stimulus with the intervals "no experience", "a weak glimpse", "an almost clear experience", and "a clear experience". For response confidence, participants were presented with the response confidence (CR) scale with the intervals "unsure", "slightly confident", "almost certain", and "certain". Participants were instructed to introspect on awareness and confidence as two distinct states, and to use the two scales to report on these states independently of one another.

Prior to starting the experiment, participants completed a thresholding task, in which the orientation of the signal Gabor patch was adapted online using a stochastic approximation algorithm [21] until responses indicated that the participant was performing at an accuracy level of about 65%. The procedure consisted of approximately 120–140 trials, depending on the rate of convergence. SOA was held constant at 64 ms throughout. This was done to ensure that the experimental task was equally difficult for all participants. No feedback was presented during this task.

In the experiment, participants performed the two-alternative forced choice task, with the orientation of the tilted patch held constant at the pre-determined 65% threshold level. SOA between the stimulus and the mask images was freed to vary across trials, and could be 64 ms, 80 ms, 96 ms, or 112 ms. Following each trial, participants were asked to rate their level of confidence in their response on that trial. Also following each trial, participants were asked to use the PAS scale to rate their degree of perceptual awareness of the tilt in the stimulus image. The order of the two different ratings was randomised between trials.

After participants had made both ratings, feedback was given on

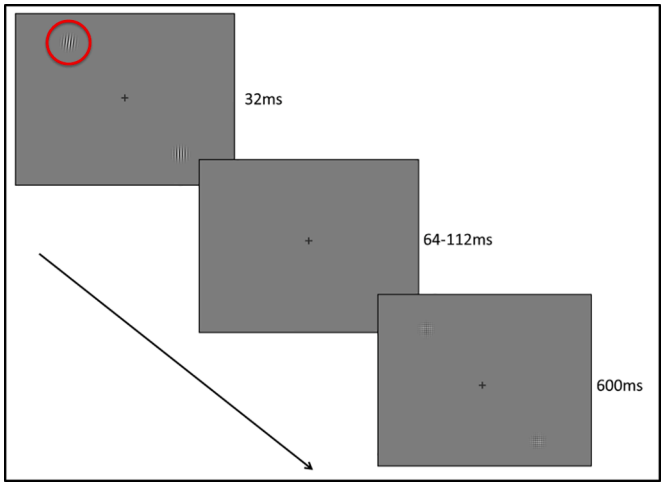

Fig. 1. Representation of the stimulus and timing used in the experiment. Participants were asked to identify the side on which the tilted Gabor patch was presented (i.e. red circle). Stimuli were presented for 32 ms, and post masked with plaids, which were presented for 600 ms. Task difficulty was varied by varying the inter-stimulus-interval, in which a blank screen was presented for either 64 ms, 80 ms, 96 ms, or 112 ms. After giving their guess on each trial, participants were asked to give separate ratings of their awareness of the stimuli, and their confidence in their responses, using 4 point scales.

whether their initial response had been correct. Feedback was given visually. For one group of participants (True Feedback group, $n = 30$), feedback truthfully reflected accuracy. For a second group of participants (False PAS feedback, $n = 29$), whenever participants provided a PAS rating of 2 (corresponding to the response "weak glimpse"), feedback indicated that their response had been incorrect even if it was accurate. For a third group of participants (False CR feedback, $n = 31$), whenever participants provided a CR rating of 2 (corresponding to the response "slightly confident"), feedback indicated that their response had been incorrect, even if it was accurate. In this way, we sought to use false feedback to directly manipulate participants' perception of the reliability of their awareness or confidence.

A block consisted of one trial for each target side (left or right) x rating order (CR first or PAS first) x ISI (48 ms, 64 ms, 80 ms, or 96 ms) combination, for a total of 16 trials per block. Order of presentation was randomised within blocks, and 40 blocks were presented, for a total of 640 trials per participant (approximately 760 trials including thresholding). The entire experiment took about 1 – 1½ h per participant to complete.

Results

For all analyses, statistical modeling was conducted using Gibbs sampling with JAGS software [14], implemented through R using the R2jags package [22]. For models investigating the main experimental results, three chains of 5000 samples each were generated, with the first 1000 samples discarded as burn-in. All chains used random initialization. For all experimental effects, we report $\widehat{R}$ metrics for convergence diagnostics, mean posterior parameter estimates, Bayesian credible intervals, and Bayes Factors calculated as the Savage-Dickey density ratio at parameter value 0. All data, model files, and analysis code are available on the Open Science Framework (https://osf.io/cfk43/).

To evaluate the thresholding procedure, we applied Bayesian linear regression to infer the effect of individual stimulus orientation on choice accuracy. This was measured as the overall hit-rate across the experiment, for all trials on which the stimulus was presented with an SOA of 64 ms. This comparison was only for the 64 ms trials, because this was the stimulus duration during the thresholding task. If thresholding was successful, then the model should support the inference that there was no relationship between threshold orientation and accuracy, and that all participants were performing the task at comparable levels of difficulty. This is important because from this, we could conclude that any differences in choice accuracy or metacognitive reports between groups were not the result of intrinsic differences in skill on the task.

To model the effect of orientation on hit-rates, we used beta regression, a commonly used method for inference with rates [4]. Individual participant hit-rates were modeled using a re-parameterised beta distribution with a probit link function. The regression model (including priors) was specified as follows:

$$\alpha \sim Normal(0, 0.1)$$

$$\beta \sim Normal(0, 0.1)$$

$$\sigma_s \sim Uniform(0, 100)$$

$$probit\left(\mu_s^{HR}\right) = \alpha + \beta.Orientation$$

$$Shape1_s = \mu_s^{HR}.\sigma_s$$

$$Shape2_s = \left(1 - \mu_s^{HR}\right).\sigma_s$$

$$HR_s \sim Beta(Shape1_s, Shape2_s)$$

where HR_s is the measured hit-rate or percent correct for a participant s, $Shape1_s$ and $Shape2_s$ are the subject level parameters for the Beta model, μ_s^{HR} and σ_s are the mean and precision of the (re-parameterised) Beta distribution, α is the linear model intercept, and β is the effect of (unstandardized) orientation on hit-rate.

$\widehat{R}$ for all parameters was < 1.02. The mean of the posterior for the model intercept (representing the probit transformed hit rate) was 0.43 (95% Credible Intervals = 0.30 to 0.58, $BF > 100$). The mean of the posterior for the effect of orientation was 0.004 (95% Credible Intervals = −0.001 to 0.009, $BF = 0.002$).

These parameter estimates suggest that although participants had higher choice accuracy during the experiment (70% correct on average, compared to the 65% thresholding target) there was no effect of individually thresholded orientation level on task performance. Thus we may conclude that thresholding was effective, and that any differences in choice accuracy between participants or groups in the experiment were not the result of individual differences in detection abilities.

To check whether participants correctly interpreted the instructions, and whether they could provide separable perceptual awareness and response confidence reports, we measured the proportion of trials in which participants gave the same PAS and CR ratings. If participants interpreted or experienced awareness to be the same as confidence, then this proportion should be close to or equal to one. The average proportion of responses for which participants gave identical PAS and CR responses was much lower (mean =0.66, sd = 0.19), suggesting that participants did interpret and use the scales distinguishably.

To test the effects of the main feedback manipulations on hit-rate, PAS ratings, and CR ratings, we applied separate beta regression models to each of these measures. We used separate models to infer the effects of false PAS feedback and false CR feedback compared to controls, and we compared the effects of the two kinds of false feedback directly to one another.

Hit-rate was modeled using a re-parameterised beta distribution with a probit link function [4]. The model included the effect of feedback (i.e. group), individual slopes for the (standardized) SOA, mean slope for SOA, and the interaction. The model (including priors) was specified as follows:

$$\alpha \sim Normal(0, 0.1)$$

$$\beta^{Group} \sim Normal(0, 0.1)$$

$$\mu_{SOA}^{\beta} \sim Normal(0, 0.1)$$

$$\beta^{Interaction} \sim Normal(0, 0.1)$$

$$\sigma_s \sim Uniform(0, 100)$$

$$\sigma^{SOA} \sim Gamma(0.01, 0.01)$$

$$\beta_s^{SOA} \sim Normal\left(\mu_{SOA}^{\beta}, \sigma^{SOA}\right)$$

$$probit\left(\mu_{s,l}^{HR}\right) = \alpha + \beta^{Group}.Group_s + \beta_s^{SOA}.SOA_{s,l} + \beta^{Interaction}.Group_s.SOA_{s,l}$$

$$Shape1_{s,l} = \mu_{s,l}^{HR}.\sigma_s$$

$$Shape2_{s,l} = \left(1 - \mu_{s,l}^{HR}\right).\sigma_s$$

$$HR_{s,l} \sim Beta\left(Shape1_{s,l}, Shape2_{s,l}\right)$$

where $HR_{s,l}$ is the measured hit-rate for participant s in SOA level l, $Shape1_{s,l}$ and $Shape2_{s,l}$ are the subject level parameters for the Beta distribution, μ_s^{HR} and σ_s are the mean and precision of the (re-parameterised) Beta distribution, α is the linear model intercept, β^{Group} is the effect of feedback group, β_s^{SOA} is the subject level effect of SOA, σ^{SOA} is the precision of the subject level effect of SOA, $\mu_{SOA}^{\beta}$is the mean effect of

SOA, and $\beta^{Interaction}$ is the interaction effect.

For the ratings, mean rating scores were calculated within each SOA condition for each participant, giving an average rating score for each level of the SOA. These averages were then rescaled from 0 to 1. The rescaled mean ratings were then modeled using a re-parameterised beta distribution with a probit link function. The group level model included the same effects as the model for hit-rate. The model was specified as follows:

$$\alpha \sim Normal(0, 0.1)$$

$$\beta^{Group} \sim Normal(0, 0.1)$$

$$\mu^{\beta}_{SOA} \sim Normal(0, 0.1)$$

$$\mu^{\beta}_{Interaction} \sim Normal(0, 0.1)$$

$$\sigma^{SOA}_{s} \sim Gamma(0.01, 0.01)$$

$$\sigma^{Interaction}_{s} \sim Gamma(0.01, 0.01)$$

$$\beta^{SOA}_{s} \sim Normal\left(\mu^{\beta}_{SOA}, \sigma\right)$$

$$\beta^{Interaction}_{s} \sim Normal\left(\mu^{\beta}_{Interaction}, \sigma^{Interaction}_{s}\right)$$

$$\sigma_s \sim Uniform(0, 100)$$

$$probit\left(\mu^{Ratings}_{s,l}\right) = \alpha + \beta^{Group}.Group_s + \beta^{SOA}_{s}.SOA_{s,l} + \beta^{Interaction}.Group_s.SOA_{s,l}$$

$$Shape1_{s,l} = \mu^{Rating}_{s,l}.\sigma_s$$

$$Shape2_{s,l} = \left(1 - \mu^{Rating}_{s,l}\right).\sigma_s$$

$$Rating_{s,l} \sim Beta\left(Shape1_{s,l}, Shape2_{s,l}\right)$$

where $Rating_{s,l}$ is the average (rescaled) rating for participant s in SOA level l, $Shape1_{s,l}$ and $Shape2_{s,l}$ are the subject/SOA level parameters for the Beta distribution, $\mu^{Rating}_{s,l}$ and σ_s are the mean and precision of the (re-parameterised) Beta distribution, α is the linear model intercept, β^{Group} is the effect of group, β^{SOA}_{s} is the participant level effect of SOA, μ^{β}_{SOA} is the mean effect of SOA, $\beta^{Interaction}_{s}$ is subject level interaction effect, $\mu^{\beta}_{Interaction}$ is the mean interaction effect, and σ^{SOA}_{s} and $\sigma^{Interaction}_{s}$ are the precisions for the subject level linear model parameters. Individual interaction effects were included in this model to improve convergence.

The main empirical trends for the feedback manipulations are presented in Fig. 2, and the main effect of SOA on each of the measures is presented in Fig. 3. Fig. 2 indicates that mean response accuracy, mean confidence ratings, and mean perceptual awareness ratings were all lower in both false feedback conditions, compared to the true feedback control. The figure also indicates that false PAS feedback was particularly effective at reducing all three measures.

The results of the models testing this interpretation for choice accuracy (i.e. hit rate) are presented in Table 1. $\widehat{R}$ for all model parameters was < 1.02. The mean of the posterior distribution for the effect of the comparison between the control group and the false PAS feedback group suggests that hit-rate dropped when people were consistently told that

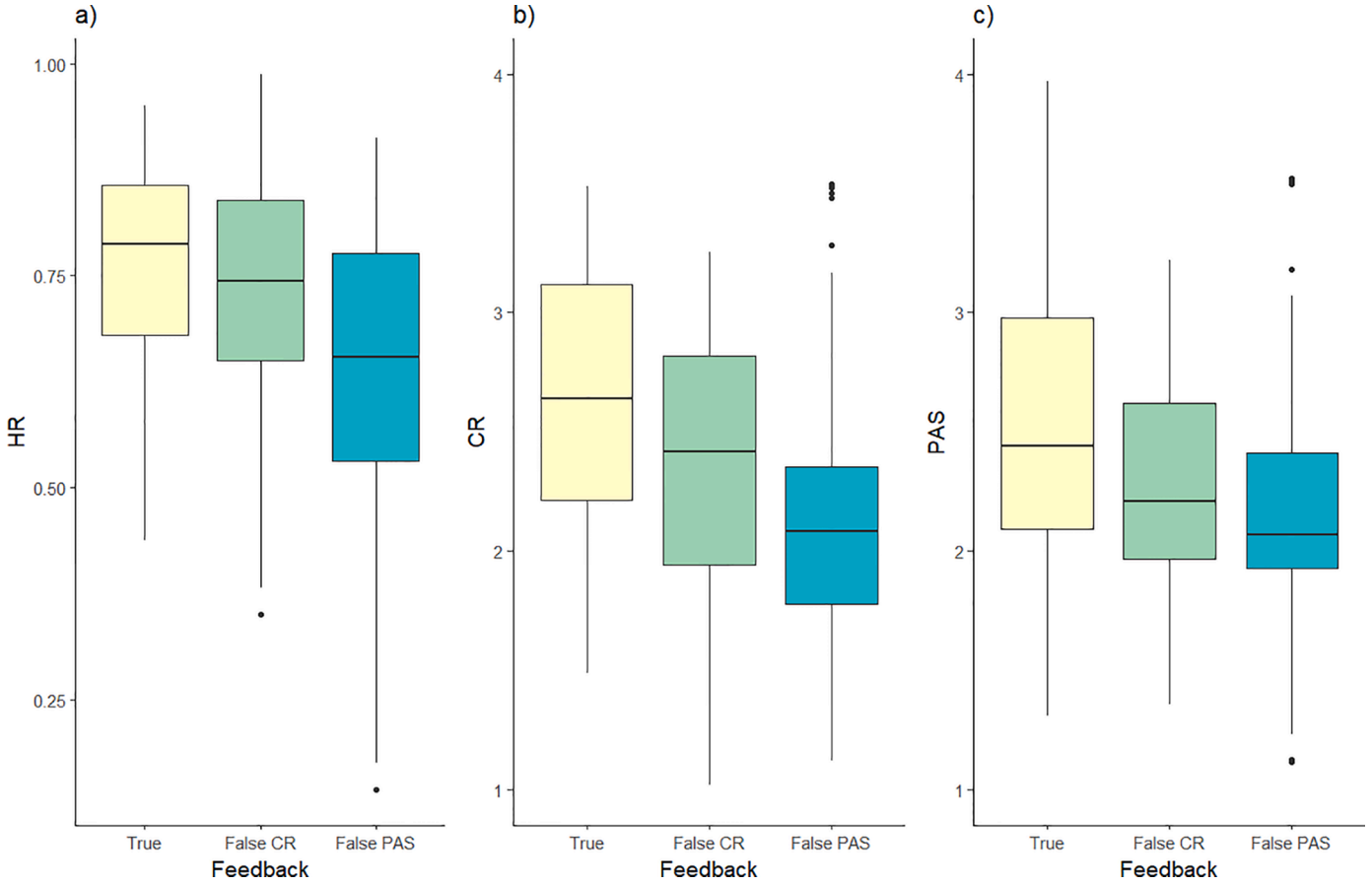

Fig. 2. Hit-rates (a), confidence ratings (b), and ratings on the Perceptual Awareness Scale (c), under conditions of true feedback, false feedback that all responses were wrong whenever participants reported that they were "uncertain" of their response, and false feedback that all responses were wrong whenever participants reported only seeing a "short-glimpse" of the stimulus. Error bars represent interquartile range, and points represent outliers.

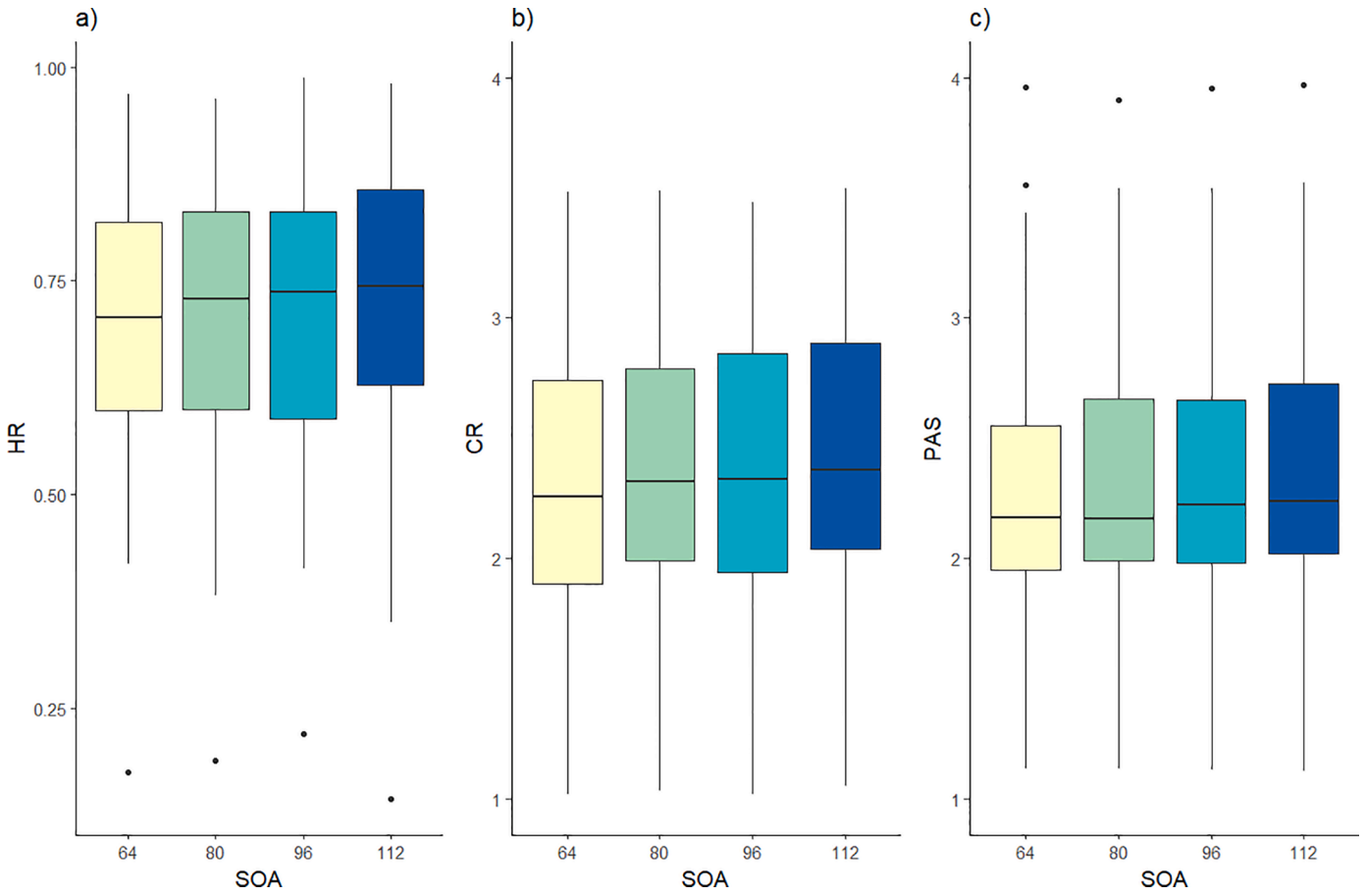

Fig. 3. Hit-rates (a), confidence ratings (b), and ratings on the Perceptual Awareness Scale (c) for different stimulus onset timings used in the experiment. Error bars represent interquartile range, and points represent outliers.

Table 1
Results of the models for the effect of false PAS feedback and false CR feedback on (probit transformed) hit-rate (compared to true feedback controls and to one another).

	Parameter Estimate (Mean of Posterior)	95% Credible Interval	Bayes Factor
False PAS vs Control			
Intercept	0.81	0.73 to 0.88	>100
Group	−0.50	−0.59 to −0.38	>100
SOA	0.04	0.01 to 0.90	0.02
Interaction	−0.01	−0.08 to 0.06	0.01
False CR vs Control			
Intercept	0.82	0.74 to 0.88	>100
Group	−0.21	−0.30 to −0.10	65.19
SOA	0.04	−0.01 to 0.09	0.03
Interaction	0.02	−0.05 to 0.09	0.01
False CR vs False PAS			
Intercept	0.61	0.55 to 0.68	>100
Group	−0.29	−0.39 to −0.20	>100
SOA	−0.04	0.01 to 0.11	0.15
Interaction	0.06	−0.10 to 0.03	0.02

all choices made on the basis of a "weak glimpse" were incorrect. The Bayesian hypothesis test provides strong evidence for this inference ($BF > 100$). The mean of the posterior distribution for the effect of the comparison between the control and the false CR feedback group suggests that people's perceptual choice accuracy also dropped when they were consistently told that all choices about which they were "slightly confident" were incorrect. The hypothesis test provides evidence for this inference ($BF = 65.19$). The mean of the posterior distribution for the effect of the comparison between the false CR and false PAS feedback groups suggests that people's perceptual choice accuracy dropped more when they received false PAS feedback. The hypothesis test provides strong evidence for this inference ($BF > 100$). Tests for all other effects provided evidence for no effect ($BF < 0.3$).

Table 2
Results of the models for the effect of false PAS feedback and false CR feedback on (probit transformed) PAS ratings (compared to true feedback controls and to one another).

	Parameter Estimate (Mean of Posterior)	95% Credible Interval	Bayes Factor
False PAS vs Control			
Intercept	0.19	0.13 to 0.25	>100
Group	−0.14	−0.21 to −0.07	10.27
SOA	0.02	−0.02 to 0.07	0.01
Interaction	−0.01	−0.07 to 0.06	0.01
False CR vs Control			
Intercept	0.16	0.09 to 0.23	>100
Group	−0.07	−0.15 to 0.008	0.06
SOA	−0.03	−0.02 to 0.07	0.01
Interaction	0.01	−0.07 to 0.05	0.01
False CR vs False PAS			
Intercept	0.06	0.02 to 0.10	0.46
Group	0.14	0.06 to 0.21	6.92
SOA	0.02	−0.02 to 0.06	0.001
Interaction	0.01	−0.06 to 0.07	0.01

The results of the models testing the effect of false feedback on PAS ratings are presented in Table 2. $\widehat{R}$ for all model parameters was < 1.03. The mean of the posterior distribution for the effect of the comparison between the control group and the false PAS feedback group suggests that people's perceptual awareness dropped when they were consistently told that all choices made on the basis of a "weak glimpse" were incorrect. The Bayesian hypothesis test provides evidence for this inference ($BF = 10.27$). The mean of the posterior distribution for the effect of the comparison between the false PAS and false CR feedback groups suggests that people's perceptual awareness also dropped relative to the false CR group. The Bayesian hypothesis test provides evidence for this inference ($BF = 6.92$). Tests for all other effects provide evidence for no effect ($BF < 0.3$).

The results of the models testing the effect of false feedback on CR ratings are presented in Table 3. $\widehat{R}$ for all model parameters was < 1.03. The mean of the posterior distribution for the effect of the comparison between the control group and the false PAS feedback group suggests that people's confidence ratings dropped when they were told that all choices made on the basis of a "weak glimpse" were incorrect. The Bayesian hypothesis test provides strong evidence for this inference ($BF > 100$). The mean of the posterior distribution for the effect of the comparison between the control and the false CR feedback group suggests that people's confidence ratings did not change when they were told that all choices about which they were "slightly confident" were incorrect. The hypothesis test provides evidence for this inference (BF = 0.03). The mean of the posterior distribution for the effect of the comparison between the false PAS and false CR feedback groups suggests that people's confidence ratings were lower the false PAS group. The Bayesian hypothesis test provides strong evidence for this inference ($BF > 100$). Tests for all other effects provided evidence for no effect ($BF < 0.3$).

Fig. 3 suggests small effects of increasing SOA on hit rate and on the two kinds of metacognitive report. However, this inference is not supported in the statistical models, and there is evidence of no interaction between the effects of feedback and SOA (all $BFs > 0.3$). This suggests that feedback effects are not dependent on task difficulty (across the ranges measured).

To determine whether the effect on performance was an effect of the feedback manipulation, we analyzed the change in response accuracy over time during the task. If the decline in accuracy was due to false feedback, accuracy should not simply be an effect of group allocation, but should rather decline over trials in the false feedback groups, as the effect of feedback exerts its influence on people's behavior.

For the analysis, we ignored SOA. We grouped all trials into 64 ten-trial bins, and calculated the hit rate within each bin for each subject. Hit rates were modeled using a re-parameterised beta distribution with a probit link function. We modeled the interaction between feedback group and trial-bin. As for the other analyses, we applied separate models contrasting the false PAS group and the false CR group to the true feedback controls, and to one another. The model was specified as follows:

$$\alpha \sim Normal(0, 0.1)$$

$$\beta^{Group} \sim Normal(0, 0.1)$$

$$\beta^{Trials} \sim Normal(0, 0.1)$$

$$\beta^{Interaction} \sim Normal(0, 0.1)$$

$$\sigma_s \sim Uniform(0, 100)$$

$$probit\left(\mu_{s,t}^{HR}\right) = \alpha + \beta^{Group}.Group_s + \beta^{Trials}.Bin_t + \beta^{Interaction}.Group_s.Bin_t$$

$$Shape1_{s,t} = \mu_{s,t}^{HR}.\sigma_s$$

$$Shape2_{s,t} = \left(1 - \mu_{s,t}^{HR}\right).\sigma_s$$

$$HR_{s,t} \sim Beta\left(Shape1_{s,t}, Shape2_{s,t}\right)$$

where $HR_{s,t}$ is the hit-rate for subject s in a ten-trial bin t, $Shape1_{s,t}$ and $Shape2_{s,t}$ are the subject/trial bin level parameters for the Beta distribution, $\mu_{s,t}^{HR}$ and σ_s are the mean and precision of the (re-parameterised) Beta distribution, α is the intercept for the linear model, and β^{Group}, β^{Trials}, $\beta^{Interaction}$ are the effects of false feedback condition or group, trial bin, and the interaction between the two.

The results of the models are presented in Table 4. $\widehat{R}$ for all model parameters was < 1.01. The mean of the posterior distribution for the effect of trial bins suggests that hit-rate increased over trials in the control conditions, but that this improvement was reduced or reversed in the false PAS and false CR groups. This interpretation is supported by Fig. 4, which plots the average hit-rate for each trial bin within each feedback condition. The figure suggests hit-rate improved slightly throughout the experiment for the control group, remained approximately constant for the false CR group, and decreased slightly for the false PAS group. This interpretation is mostly supported statistically.

Table 3
Results of the models for the effect of false PAS feedback and false CR feedback on (probit transformed) CR Ratings (compared to true feedback controls and to one another).

	Parameter Estimate (Mean of Posterior)	95% Credible Interval	Bayes Factor
False PAS vs Controls			
Intercept	0.27	0.19 to 0.35	>100
Group	−0.22	−0.31 to −0.12	>100
SOA	0.03	−0.01 to 0.07	0.02
Interaction	0.00	−0.07 to 0.07	0.02
False CR vs Controls			
Intercept	0.18	0.08 to 0.29	62.84
Group	−0.05	−0.17 to 0.05	0.03
SOA	−0.03	−0.02 to 0.08	0.02
Interaction	−0.01	−0.08 to 0.07	0.01
False CR vs False PAS			
Intercept	0.35	0.27 to 0.44	>100
Group	−0.30	−0.39 to −0.21	>100
SOA	0.04	−0.01 to 0.08	0.02
Interaction	−0.01	−0.08 to 0.06	0.01

Table 4
Results of the models for the effect of false PAS feedback and false CR feedback on (probit transformed) hit-rate over (binned) trials.

	Parameter Estimate (Mean of Posterior)	95% Credible Interval	Bayes Factor
False PAS vs Controls			
Intercept	0.62	0.59 to 0.66	>100
Group	−0.37	−0.41 to −0.31	>100
Trials	0.07	0.04 to 0.10	82.40
Interaction	−0.13	−0.16 to −0.08	>100
False CR vs Controls			
Intercept	0.62	0.59 to 0.66	>100
Group	−0.07	−0.11 to −0.02	0.23
Trials	0.07	0.04 to 0.1	203.29
Interaction	−0.01	−0.13 to −0.05	78.52
False CR vs False PAS			
Intercept	0.56	0.52 to 0.59	>100
Group	−0.30	−0.35 to −0.26	>100
Trials	−0.02	−0.05 to 0.003	0.02
Interaction	−0.03	−0.07 to 0.01	0.02

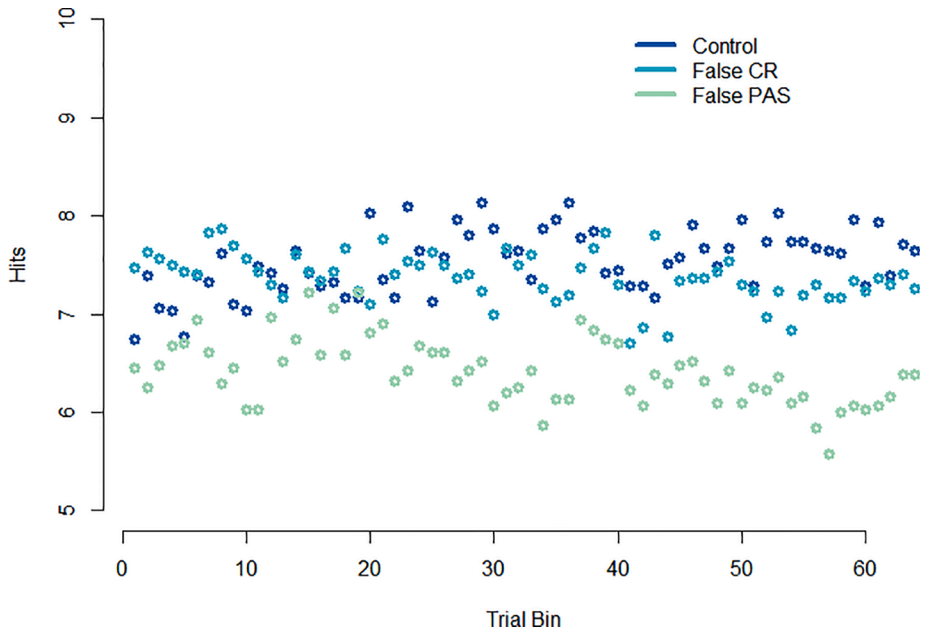

Fig. 4. Average hit-rates across trial bins for each of the feedback groups.

The Bayesian hypothesis test provides strong evidence for the inference that the false PAS feedback group's hit rate decreased over trials relative to controls ($BF > 100$ for the interaction), and for the inference that the false CR feedback group's hit rate decreased over trials ($BF = 78.52$ for the interaction), however, the same interaction was not observed in the comparison between false PAS and false CR groups ($BF < 0.3$).

To ensure that the differences observed between the false feedback groups is a specific effect of false feedback in the PAS condition, and not just some effect of differences in amount of feedback *per se* between groups, we analysed the effect of the number of instances of false feedback given during the experiment. This analysis is important, because there was a difference in the average number of false feedback events experienced by the false PAS ($M = 177.94$, $SD = 100.01$) and false CR ($M = 133.13$, $SD = 60.69$) groups. In this analysis, we directly compared the effects of false feedback on hit-rate between the two false feedback groups. The model for this analysis was specified as follows:

$$\alpha \sim Normal(0, 0.1)$$

$$\beta^{Group} \sim Normal(0, 0.1)$$

$$\beta^{Feedback} \sim Normal(0, 0.1)$$

$$\beta^{Interaction} \sim Normal(0, 0.1)$$

$$\sigma_s \sim Uniform(0, 100)$$

$$probit\left(\mu_s^{HR}\right) = \alpha + \beta^{Group}.Group_s + \beta^{Feedback}.Feedback_s + \beta^{Interaction}.Group_s.Feedback_s$$

$$Shape1_s = \mu_s^{HR}.\sigma_s$$

$$Shape2_s = \left(1 - \mu_s^{HR}\right).\sigma_s$$

$$HR_s \sim Beta(Shape1_s, Shape2_s)$$

where HR_s is the average hit-rate for subject s, μ_s^{HR} and σ_s are the parameters of the posterior distribution for the modeled subject level hit-rate, α is the intercept for the linear model, and β^{Group}, $\beta^{Feedback}$, $\beta^{Interaction}$ are the effects on hit-rate of group, number of feedback events, and the interaction between the two.

Table 5
Results of the models for the effect of the number of false feedback events on (probit transformed) hit-rate.

	Parameter Estimate (Mean of Posterior)	95% Credible Interval	Bayes Factor
Intercept	−0.67	−1.11 to −0.24	85.22
Group	0.72	0.41 to 1.05	>100
Feedback	<0.01	0.001 to 0.006	0.03
Interaction	<0.01	−0.004 to −0.001	<0.01

The results of the model are presented in Table 5. $\widehat{R}$ for all model parameters was < 1.1. The model provides strong evidence against a hypothesis that differences in the quantity of false feedback are responsible for the effects of false PAS and false HR feedback on hit-rate.

Discussion

The results of the experiment support the view that reports on perceptual awareness and reports on response confidence are different, and that participants understood their instructions on how to use the two scales distinctly.

As all interpretations of experimental findings, some limitations apply. The difficulty of the experiment was naturally determined in part by our choice of SOA range - and a different range might theoretically have yielded different results. Furthermore, our contrast between PAS and CR indicate categorically different metacognitive processes, yet using more types of report might have revealed a gradually changing pattern with PAS and CR as end poles.

We found that false feedback about PAS ratings reduced choice accuracy on the task. When participants were consistently told that their responses were wrong whenever they made a judgement about a stimulus that they had only glimpsed, they came to produce more errors. We found similar effects of this intervention on PAS ratings and CR ratings. The pattern of results was different for false CR feedback. When participants were consistently told that their responses were wrong whenever they reported that they were "slightly confident" about a response, they also produced more errors. However, the effect was weaker than for the effect of false PAS feedback, and there was evidence for no effects of false CR feedback on PAS or CR ratings.

From the experiment we may conclude that false feedback on PAS ratings – being told that one is incorrect for whatever choice one makes on the basis of a brief glimpse of the stimuli – leads to more strongly reduced choice accuracy on the task. The same intervention also leads to reduced perceptual awareness measured in PAS ratings, and reduced confidence measured in CR ratings. These results suggest that perceptual decision-making is more malleable to interventions on conscious awareness than similar interventions on metacognitive confidence. This, again, suggests that the two types of reports represent different kinds of metacognitive states. This proposal is in and of itself of relevance to a broad range of current neuroscientific research including research in psychiatry (e.g. [2,20]) and neurorehabilitation (e.g. [8]) where metacognition is typically considered as one unified type of mental states.

The finding is in line with some previous studies showing that confidence ratings are influenced by other factors than the underlying correctness itself as shown in e.g. working memory research [3]. Such frequently found weak correlations between confidence and performance are in contrast with very strong contrasts between PAS and performance as reported in many experiments (e.g. [9]). In a few previous experiments, PAS and confidence ratings have been compared directly, and those studies have all indicated that the two types of scales are used differently (e.g. Rausch & Zehetleiner, 2016; [18]). Using feedback manipulations, however, this experiment presents more direct evidence that the scales relate to different types of metacognition.

Two alternative explanations may be provided for this pattern of results. The first is that the result is not an effect of the false feedback, but is rather due to serendipitous differences in sensitivity between the different groups of participants. Importantly, participants completed an adaptive thresholding procedure prior to the main experiment. Although performance during the test was slightly higher than at thresholding (indicating some overall learning during the task), we found no relationship between individual orientation thresholds and task performance. Thus we conclude that the groups were comparable prior to the test. In addition, our result also suggest that the difference between the false PAS and control groups increases across trials during the experiment, suggesting a causal effect of the feedback manipulation.

The second alternative explanation is that the difference in performance between groups may be a causal effect of the feedback manipulation, but that this effect is not an effect of the false feedback intervention on conscious processing, and that it is rather an effect of amount of false feedback *per se*. This explanation might be motivated by a concern that participants in the false PAS group simply received more false negative feedback than participants in the CR group. Such a concern is reasonable, because we did find that people incidentally used PAS 2 more frequently than the equivalent CR rating, and so the false PAS group did receive more false feedback than the false PAS group. However, we also found evidence that differences in accuracy were not dependent on differences in the amount of false feedback given. Thus we may conclude that the effects reported here are specific to the qualitative effects of feedback on processes related to the rating scales used to make the metacognitive reports.

A dominant account of perceptual confidence is that it is directly determined by signal strength (e.g. [1]). However, if an intervention based on false feedback modifies both the report for the higher-order *and* the first-order state, this view may need to be revised. In support of a more complex view, Fleming et al. [6] found that TMS used to manipulate premotor cortex influenced confidence ratings, suggesting that at least this particular type of higher-order report relates to motor processes, as well as perceptual processes.

We are at present unable to specify the exact nature of the relation between first and higher order states. However, our results point in certain directions. Our results go against linear processing models, suggesting the idea that metacognitive states do not penetrate first order states. The results thus suggest some kind of reciprocal or a parallel relationship between first and higher order states [12], with changes in higher order states influencing how sensory evidence is used in perceptual decision-making.

Declaration of Competing Interest

The authors declare no conflict of interest for this work.

CRediT authorship contribution statement

Joshua Skewes: Conceptualization, Formal analysis, Writing – original draft. **Chhris Frith:** Conceptualization, Visualization, Writing – original draft. **Morten Overgaard:** Conceptualization, Visulization, Writing – original draft.

References

[1] S. Barthelmé, P. Mamassian, Flexible mechanisms underlie the evaluation of visual confidence, Proc. Natl. Acad. Sci. 107 (48) (2010) 20834–20839.
[2] V. Bliksted, E. Samuelsen, K. Sandberg, B. Bibby, M. Overgaard, Discriminating between first and second order cognition in first-episode paranoid schizophrenia, Cognit. Neuropsychiatry 22 (2) (2017) 95–107.
[3] S. Bona, J. Silvanto, Accuracy and confidence of visual short-term memory do not go hand-in-hand: behavioral and neural dissociations, PLoS One 9 (3) (2014).
[4] S. Ferrari, F. Cribari-Neto, Beta regression for modelling rates and proportions, J. Appl. Stat. 31 (7) (2004) 799–815.
[5] S. Fleming, C.D. Frith, The Cognitive Neuroscience of Metacognition, Springer, 2014.
[6] S.M. Fleming, B. Maniscalco, Y. Ko, N. Amendi, T. Ro, H. Lau, Action-specific disruption of perceptual confidence, Psychol. Sci. 26 (1) (2015) 89–98.
[7] R.R. Hassin, Yes it can – on the functional abilities on the human unconscious, Perspect. Psychol. Sci. 8 (2) (2013) 195–207.
[8] J. Lindeløv, R. Overgaard, M. Overgaard, Improving working memory function in brain-injured patients using hypnotic suggestion, Brain 140 (4) (2017) 1100–1106.
[9] M. Lohse, M. Overgaard, Emotional priming depends on the degree of conscious experience, Neuropsychologia 128 (2019) 96–102.
[10] J. Metcalfe, A.P. Shimamura, Metacognition: Knowing about Knowing, MIT Press, Cambridge, 1994.
[11] E. Norman, M. Price, Measuring consciousness with confidence ratings, in: M. Overgaard (Ed.), Behavioral Measures in Consciousness Research, Oxford University Press, 2015.
[12] M. Overgaard, J. Mogensen, An integrative view on consciousness and introspection, Rev. Philos. Psychol. 8 (2017) 129–141.
[13] M. Overgaard, K. Sandberg, Kinds of access: different methods for report reveal different kinds of metacognitive access, Philos. Trans. R. Soc. Lond. B Biol. Sci. 367 (2012) 1287–1296.
[14] M. Plummer, JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling, in: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, March 20–22, Vienna, Austria, 2003.
[15] A. Pouget, J. Drugowitsch, A. Kepacs, Confidence and certainty: distinct probabilistic quantities for different goals, Nat. Neurosci. 19 (2016) 366–374.
[16] T.Z. Ramsøy, M. Overgaard, Introspection and subliminal perception, Phenomenol. Cognit. Sci. 3 (1) (2004) 1–23.
[17] M. Rausch, S. Hellmann, M. Zehetleitner, Confidence in masked orientation judgments is informed by both evidence and visibility, Atten. Percept. Psychophys. 80 (2018) 134–154.
[18] K. Sandberg, B. Timmermans, M. Overgaard, A. Cleeremans, Measuring consciousness: is one measure better than the other? Conscious. Cognit. 19 (2010) 1069–1078.
[19] in: K. Sandberg, M. Overgaard, Using the perceptual awareness scale (ed), in: M. Overgaard (Ed.), Behavioral Methods in Consciousness Research, Oxford University Press, 2015.
[20] X. Sun, C. Zhu, S. So, Dysfunctional metacognition across psychopathologies: a meta-analytic framework, Eur. Psychiatry 45 (2017) 139–153.
[21] B. Treutwin, Adaptive psychophysical procedures, Vis. Res. 35 (17) (1995) 2503–2522.
[22] Yu-Sung, S., & Yajima, M. (2012). R2jags: a Package for Running jags from R. *R package version 0.03-08, URL http://CRAN. R-project. org/package= R2jags*.
[23] M. Zehetleitner, M. Rausch, Being confident without seeing: what subjective measures of consciousness are about? Atten. Percept. Psychophys. 75 (2013) 1406–1426.